

You Trust Your Chatbot With Everything. Should You?

PART 2

*Governments, Courts, and the Battle
Over Your Chatbot Conversations*



Theodore Christakis





Theodore Christakis is Professor of International, European and Digital Law at University Grenoble Alpes (France), Director of Research for Europe with the Cross-Border Data Forum, Member of the Board of Directors of the Future of Privacy Forum and a former Distinguished Visiting Fellow at the New York University Cybersecurity Centre. He is co-Chair with Héber Arcolezi on “Responsible AI: Design, Regulation and Conformity” at the Multidisciplinary Institute in Artificial Intelligence.



Cover Illustration:

Original illustration conceived and directed by the author and produced with generative AI assistance.

Artistically inspired by the surrealist aesthetic associated with René Magritte.

Reproduction permitted for non-commercial purposes with full attribution to the author and citation of the study. © 2026 Theodore Christakis.

To cite this study: T. Christakis, *You Trust Your Chatbot With Everything. Should You? Part 2: Governments, Courts, and the Battle Over Your Chatbot Conversations*, AI Regulation Papers, 26-06-1, [AI-Regulation.com](https://ai-regulation.com), June 2026.

These statements are attributable only to the author, and their publication here does not necessarily reflect the view of the other members of the AI-Regulation Chair or any partner organisations.

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-23-IACL-0006) and by the Interdisciplinary Project on Privacy (IPoP) of the Cybersecurity PEPR (ANR 22-PECY-0002 IPOP).

Preface and Acknowledgements

For a year now I have been caught up in a slightly crazy project: to map, as precisely as I could, what becomes of privacy once people start confiding in AI chatbots.

Two things set it in motion. The first was personal. In conversations with the people around me, family, business colleagues, friends, it became clear that a great many of them, myself included, now turn to chatbots for all sorts of things, and often for very private ones, beginning with advice about their own health or the health of the people they love. That pushed me towards a simple question with a complicated answer: how well are these intimate exchanges protected in practice? When, and for what reasons, can the providers themselves read and use them? When, and for what reasons, can they be reached by people outside the providers, whether law enforcement, intelligence agencies, an opposing party in litigation, or an attacker? How secure are they? And how far could privacy by design, together with sound management choices, strengthen their protection at a moment when both the number of users and the intimacy of what they confide are rising spectacularly?

The second was the realisation of how little there was to build on. When I started, the questions I have just described had drawn neither much regulatory attention nor sustained academic study. On the regulatory side, most of the guidance issued so far on generative AI, in Europe at least, has concerned web scraping and the training of large language models on data already public on the internet. That is an important question, but it is a different one, and it says little about what passes through the millions of intimate exchanges taking place every day, a great deal of which involves sensitive data. OpenAI alone has said that some 230 million people a week, and more than 40 million a day, bring it questions about their health. On the academic side, there was simply no study that tried to take the subject as a whole; the valuable work that exists tends to treat single issues, and most of it appeared only in the last few months, well after I had begun. I have tried, so far as I could, to take all of it into account.

I will not pretend the work was easy. More than once it felt less like research than like a punishment out of Greek myth. Some weeks it was the boulder of Sisyphus, pushed to the top of the hill only to roll back down. Others it was the cask of the Danaids, the vessel those sisters were condemned to fill for ever although it could never hold water, so that the task was never done. Most often it was the Hydra of Lerna: each question I cut down grew two more in its place, every one as important as the last.

What I publish today is the most ambitious part of a series. It opened with [Part 1, How the Controller Uses Your Chat Data](#), and a [companion article for the IAPP](#), which mapped what I called the internal boundary: how the providers themselves handle what we tell them, from training and human review to advertising and the sharing of data across their own services, across five of the main consumer chatbots. Its central proposal was a Sealed Mode for the most sensitive conversations, a pathway whose protection would rest on architecture rather than on promises. Two shorter studies followed as the ground kept shifting. One, on the [rush of health AI agents](#), with its own [IAPP companion](#), and a dedicated [op-ed in Le Monde](#), looked at how, in barely three months, five companies wired electronic health records, wearables and wellness apps into their chatbots and turned them into personal health data hubs, and at the governance questions that speed left behind. The other, [written with Peter Swire](#), read Meta's announcement of 13 May 2026 as the first real-world appearance of exactly that Sealed Mode architecture: a conversation the provider has arranged, by design rather than by policy, to be unable to read.

What I publish today turns to the other side of the boundary, the external one, and it is the hardest of the series. It follows a single retained conversation into the four places where it can pass out of the user's control: disclosure by the provider on its own initiative, access compelled by governments, discovery in litigation, and exposure through cyberattacks and breaches. It then asks what architecture and habit, rather than legal right alone, could do to protect it.

It was difficult for two reasons. The first is that each of the four chapters turns on legal questions that are both hard and far from settled. To take only two: in a copyright case brought against OpenAI, a

court ordered the company to preserve and hand over the ChatGPT conversations of people who had nothing to do with the litigation, an order that swelled within months to cover more than a hundred million conversations; and two federal courts, within weeks of one another, split on whether typing your thoughts into an AI tool strips them of the protection they would have kept on paper, leaving wide open the question of whether there is, or should be, anything like an "AI privilege."

The second reason is that most of this fast-moving law is American. Europeans often assume that AI is barely regulated in the United States. This study suggests something close to the opposite: that some of the most important questions about the privacy of these conversations are being decided right now, in American courtrooms, and that the answers will shape the field for everyone. I am a European lawyer, not an American one. I have done my best to research and present these questions as accurately as I could, and I have had the rare good fortune of help from some of the finest US lawyers working on them.

I am profoundly grateful to Richard Salgado and Peter Swire, who read earlier drafts with extraordinary care and sent successive waves of comments that improved this work beyond anything I could have managed on my own. If it is far stronger than the draft they first saw, the credit is theirs. I warmly thank Jules Polonetsky and Deborah Kennedy-Mayo for their excellent comments on particular and unsettled points of US law. For their reading and feedback on important questions of EU law and on technical matters, I am equally grateful to Yann Padova, T. Karathanasis, Mikołaj Barczentewicz and Claude Castelluccia. I learned a great deal, too, from my fellow panellists in Venice, at the Privacy Symposium, on the panel "[Whispers to Machines: When Chatbots Become Evidence](#)", and at CPDP in Brussels, on "[My Chatbot, My Confidant?](#)"; I want to thank in particular William Malcolm, Idriss Kechida, Declan McDowell-Naylor and Natascha Gerlach, who made up these panels with Peter Swire and Yann Padova. Last but not least, I am grateful to the regulators, and to the many people in industry, including several at the companies examined in this series who provided factual feedback, for taking the time to talk these questions through with me.

A clear word is owed in return. All of them, my US colleagues first among them, gave the best of their knowledge and their time. The errors that remain, and in a field moving this quickly there are bound to be some, are mine and mine alone. I hope the reader will forgive them in view of the sheer size of the monster I was trying to describe.

As a non-native English speaker, I used generative AI tools (Claude, ChatGPT and Gemini) for substantial help with drafting, rewriting and editing the text. I also used these tools for research assistance, to help prepare the tables, to challenge and discuss my arguments, to check their accuracy and to identify points I had missed. The ideas, analysis, arguments and recommendations are my own. I have read and verified every source cited, and I take full responsibility for the content.

And a word about that size. I hope no one will be put off by the length of what follows. Mapping at once the scale of what is happening and the legal questions it raises could not, I am afraid, be done briefly.

If this study does its work, it will not be the last word on any of these questions. I would rather it be a point of departure: a map that others can test, correct and extend. A great deal of work remains, and my main hope is that this makes some of it a little easier.

Table of Contents

EXECUTIVE SUMMARY	1
Introduction	4
1. Proactive Corporate Disclosure of Consumer Chatbot Logs.....	9
1.1. The legal landscape: existing obligations and what is distinctive about the chatbot context	9
1.2. When moderation becomes referral: three recent episodes	15
1.3. OpenAI after the Canadian tragedy	19
1.4. The mounting legal pressure on providers, and its effect on referral incentives	21
1.5. Child safety as the most standardised domain of proactive reporting.....	22
1.6. The tension with cryptographic hardening and Sealed Mode	24
1.7. Lessons from social media moderation, Google search, and human confidants: why chatbots raise distinctive accountability questions	27
1.8. Conclusions and tentative recommendations.....	34
Tentative Recommendations on Proactive Corporate Disclosure.....	36
2. When the State Compels Disclosure: Law Enforcement, Intelligence Agencies, and the Surveillance Multiplier	39
2.1. The basic legal ladder	40
2.2. User notice, resistance, and cross-border process.....	41
2.3. Preservation and the practical defeat of ephemerality	43
2.4. Reverse prompt warrants: when the state searches for the user, not the data	44
2.5. The foreign intelligence dimension: Could Section 702 Be Used to Compel Reverse-Prompt Searches?	47
2.6. Memory, agentic AI, and the intelligence “gold mine”	50
2.7. Defence contracts and consumer data: does the dual role expose consumer chats?	51
2.8. Transparency reports: valuable but incomplete	52
2.9. Conclusions and tentative recommendations.....	55
Tentative Recommendations on Compelled Government Access	57
3. Discovery: When Chatbot Conversations Become Evidence	59
3.1. What exactly happened in the New York Times discovery battle?	60
3.2. OpenAI’s privacy defences and why they failed.....	62
3.2.1. <i>The substance of OpenAI’s privacy defence</i>	62
3.2.2. <i>Why the defence did not prevail</i>	63
3.3. “AI privilege”	64
3.3.1. <i>The first question: no privilege shields a provider’s records of consumer AI conversations</i> ..	64
3.3.2. <i>Sam Altman’s proposal, and the trouble with a privilege the provider would hold</i>	64
3.3.3. <i>The second question: does using an AI tool forfeit the user’s own protection? (Heppner and Warner)</i>	65
3.3.4. <i>Why the Heppner reasoning deserves critical scrutiny</i>	67
3.3.5. <i>Conclusion: both questions turn on architecture</i>	69
3.4. Why de-identification could not fully solve the problem	70
3.5. Why protective orders and judicial safeguards do not eliminate confidentiality risk.....	70
3.6. Could a NYT-style discovery order happen in Europe?	72
3.7. Beyond the New York Times: from bulk discovery to the everyday litigant.....	75
3.7.1. <i>Conversations sought from the provider</i>	75
3.7.2. <i>Conversations sought from the user</i>	76
3.8. Why similar discovery fights may persist and multiply	77
3.9. Conclusions and tentative recommendations.....	77
Tentative Recommendations on Discovery and Evidence	79
4. Cyberthreats, Insider Risks and Other Data Breaches	82
4.1. Why chatbot logs are a distinctive breach target.....	82

4.2. The public record: warning incidents, not yet a full catastrophe	84
4.2.1. Infrastructure vulnerabilities and software bugs	84
4.2.2. Sharing design failures and accidental publication	85
4.2.3. Client-side and supply-chain exfiltration	85
4.2.4. Tracker and telemetry leakage from providers' own platforms	86
4.2.5. Credential compromise and enterprise leakage	88
4.3. The new risk introduced by agents	89
4.4. Insider access and vendor exposure	90
4.5. The legal lens: breach notification is necessary but not enough	90
4.6. Conclusions and tentative recommendations	91
Tentative Recommendations on Cyberthreats, Insider Risks and Data Breaches	92
5. Conclusions and Consolidated Recommendations	93
5.1. The four pathways and what they share	93
5.2. Building on existing mechanisms, addressing what is distinctively new	94
5.3. The architectural turn: Meta's 13 May 2026 announcement	94
5.4. Open questions	95
5.5. Consolidated Tentative Recommendations	97

List of Tables

Table 1. Comparative overview of proactive corporate disclosure	33
Table 2. Comparative overview of compelled government access	42
Table 3. Transparency reporting and disclosure metrics	55

EXECUTIVE SUMMARY

You Trust Your Chatbot With Everything. Should You?

Part 2: Governments, Courts, and the Battle Over Your Chatbot Logs

Hundreds of millions of people now confide in AI chatbots as though the conversation were private. In law, it is not. This study follows those conversations into the four places they can be exposed, and asks what should be done.

In December 2025 a federal court in Manhattan ordered OpenAI to hand twenty million ChatGPT conversations to opposing counsel in a copyright case. Within three months the demand had expanded to a further 108 million de-identified conversations in all. They belonged to ordinary people who sometimes had typed their symptoms, their debts, their fears and their secrets into a system they treated as private.

That behaviour is now ordinary. People often speak to consumer chatbots in the register once reserved for a doctor, a lawyer, a diary, or no one at all; the interfaces are built to invite it; and the providers themselves describe the result as “among the most sensitive information in your digital life”. Yet in law a chatbot conversation is an ordinary record. It carries none of the protections that attach to the same words in a consulting room or a lawyer’s office. What shields it is not a legal right but the provider’s retention choices and the user’s own habits, and both can fail. This study, the second part of a project on consumer-AI privacy, is about that gap, and about the routes by which a conversation leaves the user’s control.

Its central finding is a convergence. The very choices that make a chatbot useful, namely retention, memory, logging, personalisation and connected tools, are the choices that make the resulting record preservable, searchable, discoverable, disclosable and exploitable. A single stored conversation is, at the same moment, a candidate for police referral, a target for a government demand, evidence in a lawsuit, and an asset in a data breach. No prior work, to the author’s knowledge, has examined these four exposures together.

Four pathways out of the confidential space

The provider refers it to the police. Companies decide, largely in private and at scale, whether an alarming conversation should be sent to law enforcement. Three episodes from 2025 and 2026, a mass shooting in Tumbler Ridge, the Florida State University shooting, and an armed police call-out in Strasbourg, show the stakes. The governing rule is the decades-old emergency-disclosure framework built for email and messaging, now applied to material that is unusually intimate, hard to interpret, and judged by automated classifiers with human review on escalation. The criteria, the volumes and the error rates are almost entirely outside public view.

The state compels it. Subpoenas, court orders and warrants reach chatbot data as they reach any provider’s records, and preservation requests quietly defeat the “delete” button and the temporary-chat settings users rely on. The new frontier is the reverse-prompt warrant, which asks not about a known suspect but for anyone who entered a given prompt; the Supreme Court’s geofence case, *Chatrie*, shows how near that prospect now is. Memory and agentic features turn a transcript into a longitudinal profile and a record of real-world actions, multiplying what a single demand can reveal.

A litigant obtains it in discovery. The *New York Times* case shows conversations produced by the million, overwhelmingly those of strangers to the suit, with no “AI privilege” to stop it. The study separates two questions the debate runs together: there is no privilege over the provider’s records, and it remains unsettled, after the conflicting *Heppner* and *Warner* rulings, whether using an AI tool strips a litigant’s own preparation of work-product protection. The exposure is not confined to famous cases. An ordinary person’s conversations can be demanded in their own divorce, employment or business dispute, and European procedure reaches that same everyday route.

A breach exposes it. Stored chatbot conversations are a concentrated, high-value target, and the public record already holds a run of warning incidents, from open databases and sharing-feature leaks to

tracker and credential exposure. Agentic AI, able to read mail, files and accounts, widens the attack surface again. Breach-notification law is necessary but does not reach the prior question, which is how much sensitive data is retained, and in what form, before anything goes wrong.

The architectural turn

As the study went to press, Meta launched, on 13 May 2026, a WhatsApp mode in which it cannot read the user's conversation with Meta AI, because the chat is processed inside hardware the provider has been built unable to see into. This is the first mass-market instance of confidentiality enforced by design rather than by promise, and it vindicates the architectural concept, **Sealed Mode**, that Part 1 of the project proposed. It also reframes every pathway above: a conversation the provider cannot read cannot easily be referred, compelled, discovered or breached. But it opens a hard question in the other direction. A provider that cannot see a conversation cannot step in to prevent harm, which reopens the long encryption debate in a setting where the provider is itself one party to the talk.

What should be done. The study's stance is constructive. The existing legal and institutional machinery is not obsolete; the chatbot setting inherits it and needs targeted refinement rather than wholesale replacement. Real protection takes two layers working together: policy, which here means not only law and contracts but the management and operational choices made by product, engineering and security teams, and architecture, which can supply an assurance that policy alone cannot. On that basis the study offers nineteen recommendations, addressed to providers, courts, regulators and legislators on both sides of the Atlantic and summarised overleaf. It is written by a European lawyer as an informed external observer, is grounded mainly in US law, where the leading cases and the fullest public record currently sit, and is offered, throughout, in the spirit of opening a debate rather than closing it.

The nineteen recommendations at a glance

Each is set out in full at the end of its chapter; the wording below is a plain-language précis. The four colours track the four pathways.

#	RECOMMENDATION, IN BRIEF
PROACTIVE CORPORATE DISCLOSURE (CHAPTER 1)	
1	Publish aggregate figures on provider-initiated police referrals: how many, and on what grounds.
2	Audit the criteria that decide which conversations enter the referral pipeline.
3	Separate three kinds of retention: legal preservation, user-chosen storage, and provider-kept data.
4	Calibrate a chatbot's encryption to the environment it operates in.
5	Tell users plainly what kind of relationship a chatbot actually is.
COMPELLED GOVERNMENT ACCESS (CHAPTER 2)	
6	Report government demands for chatbot data, including the new reverse-prompt and memory demands.
7	State publicly that bulk "who entered this prompt" searches will be resisted.
8	Treat memory profiles and agent action logs as distinct categories in legal process.
DISCOVERY AND EVIDENCE (CHAPTER 3)	
9	Design retention, memory and Sealed Mode with litigation exposure in mind.
10	Build the architecture on which a limited "AI privilege" could rest.

#	RECOMMENDATION, IN BRIEF
11	Resolve whether using an AI tool forfeits work-product protection (the Heppner/Warner split).
12	Develop proportionality guidance for discovery aimed at millions of non-parties' conversations.
13	Strengthen de-identification and govern derived datasets, excerpts and embeddings.
14	Warn users their own chats are discoverable in their own disputes, and give real deletion controls.
CYBERTHREATS, INSIDER RISKS AND DATA BREACHES (CHAPTER 4)	
15	Classify chatbot conversations as high-sensitivity data by default.
16	Shrink the breach blast radius: retain less, segment, and use Sealed Mode as a security control.
17	Apply least-privilege access to insiders, vendors and AI agents.
18	Make sharing private by default; judge breaches by substance, not only identifiers.
19	Publish security-incident and breach transparency, not only government-request reports.

Introduction

In December 2025, a federal magistrate judge in Manhattan ordered OpenAI to produce twenty million ChatGPT conversations to opposing counsel in a copyright case. Three months later, the production fight expanded to a newly identified repository of 88 million more, all de-identified. They belonged to ordinary people, strangers to the litigation, who had spent intimate and sometimes desperate hours typing into a system they experienced as private. Around the same time, another federal judge in the same district held that documents a criminal defendant had generated with Claude, recovered from his own devices, were protected by neither attorney-client privilege nor the work-product doctrine. Six weeks later, the United States Supreme Court heard argument in *Chatrie*, a case about a geofence warrant that had compelled a provider to identify every user whose device was recorded in a given area, a technique the Brennan Center warned could, if extended to generative AI, support warrants that identify “everyone who issued a prompt to an AI chatbot”.¹ None of these episodes happened in isolation. They are pieces of a single story: what happens to a chatbot conversation after the user closes the window.

Hundreds of millions of people now confide in consumer AI chatbots about health, money, family, sex, work, legal exposure, political doubts, and fears for the present and the future. The interfaces are built to invite that disclosure, and users behave as though they are in a private space. The legal rules, the corporate policies, and the technical architectures that actually govern those conversations are not always aligned with that expectation. This study is about the gap between the two.

That users experience these conversations as private is not merely an inference from interface design; it is how the providers themselves describe the material. Resisting the New York Times’ demand for twenty million ChatGPT conversations, OpenAI publicly described ChatGPT as a place where hundreds of millions of people bring deeply personal concerns, and said it treats the resulting data as “among the most sensitive information in your digital life”.² Its chief executive has gone further, arguing that talking to an AI should be like talking to a lawyer or a doctor, and calling for the recognition of an “AI privilege”.³ These are the words of the data controller itself, offered while resisting disclosure, and they sit uneasily with any view that would treat these conversations as ordinary records of an online service.

The usage research bears this out, including the providers’ own. OpenAI’s large-scale study with the National Bureau of Economic Research found that non-work messages had grown to more than 70% of consumer use by mid-2025, with practical guidance, information-seeking and writing together accounting for close to 80% of conversations.⁴ Even the comparatively narrow category of relationships and personal reflection represented around 1.9% of messages, which on the reported volumes amounts to roughly 342 million such conversations every week.⁵ Anthropic’s parallel study of Claude reported that affective conversations made up 2.9% of interactions, a small share that still covers very large

¹ Brief of Amici Curiae the Brennan Center for Justice, the Center for Democracy and Technology, the Surveillance Technology Oversight Project, and Defending Rights & Dissent in Support of Petitioner, *Chatrie v. United States*, No. 25–112 (US Sup. Ct. February 2026), discussed in Brennan Center for Justice, [Okello Chatrie v. United States of America](#).

² OpenAI (Dane Stuckey), [Fighting the New York Times’ invasion of user privacy](#) (12 November 2025); the same statement reports that around 800 million people use ChatGPT each week. See also OpenAI, [How we’re responding to The New York Times’ data demands in order to protect user privacy](#) (5 June 2025).

³ Sam Altman, “AI privilege” post on X (26 June 2025), linked in OpenAI, [Reporting the facts about the New York Times’ lawsuit](#); Altman argued that talking to an AI should be treated like talking to a lawyer or a doctor. See further Section 3.3.2 below.

⁴ Aaron Chatterji, Thomas Cunningham, David J. Deming and others, [How People Use ChatGPT](#), NBER Working Paper No. 34255 (September 2025), finding that non-work messages rose from 53% to more than 70% of consumer use, and that practical guidance, information-seeking and writing together account for nearly 80% of conversations.

⁵ American Psychological Association, [OpenAI publishes new study on how people are using ChatGPT](#) (30 September 2025), noting that relationships and personal reflection represent about 1.9% of messages, equivalent to roughly 342 million conversations a week at OpenAI’s reported volume of some 18 billion weekly messages.

absolute numbers and ranges across loneliness, relationships, career crises and existential questions.⁶ A Washington Post analysis of tens of thousands of conversations exposed through ChatGPT's sharing feature found that users routinely supplied details they would not enter into a conventional search engine.⁷ The sensitivity of chatbot data is therefore not confined to a niche of users seeking companionship; it runs through the ordinary, predominantly practical use that makes up the bulk of the traffic, in which health symptoms, financial circumstances, legal exposure and family conflict surface as a matter of course.

It might be objected that none of this is unique to chatbots, and the objection is partly right. Email, private messaging, search histories and cloud-stored documents can each be intensely revealing, and it would be neither accurate nor necessary to claim that chatbot conversations are categorically more intimate than every other form of electronic communication, still less that those other forms deserve weaker protection. The argument here is narrower. A behavioural literature on self-disclosure finds that the perceived non-judgement and anonymity of a conversational agent can make people disclose sensitive information, in some settings to a degree they would not match with another human.⁸ What is distinctive about the consumer chatbot is that this disclosure-inviting quality now coincides, inside a single continuously growing and identity-linked record held by one provider, with a register the providers themselves liken to the lawyer's office or the consulting room, with material that mixes genuine intent, fiction, role-play and venting in ways that resist confident reading, and with a scale measured in hundreds of millions of users and billions of weekly conversations. It is that concentration, rather than any single feature on its own, that marks a new dimension of privacy risk, and it is why this study follows these conversations into each of the settings in which they can leave the user's control.

This is the second part of a research project on the privacy of consumer AI chatbots. Part 1, *You Trust Your Chatbot With Everything. Should You? Part 1: How the Controller Uses Your Chat Data*, published in March 2026, compared the consumer-facing privacy protections of the five most prominent platforms in this category: ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), Grok (xAI) and DeepSeek.⁹ It examined four established boundaries of confidentiality defined by the user's relationship with the provider, namely training, human review, advertising and operational sharing, and identified a fifth, emerging boundary: memory. Memory, persistent personalisation and the integration of conversation data into longitudinal user profiles turn the chatbot from a stateless tool into something closer to a structured, provider-curated record of a person's preferences, vulnerabilities, recurring concerns and evolving private life. Part 1 also proposed an architectural concept, **Sealed Mode**, as one response: a category of conversation in which the provider verifiably narrows its own access, retention, training use, advertising signals and reuse from the moment of collection.

A follow-up study published a few weeks later, *The Health AI Agent Rush: Five Companies, Your Health Data, and the Governance Questions Nobody Is Asking*, examined how the same five companies are deploying agentic AI in healthcare and the governance questions raised by the explicit collection and processing of health data through these new agents.¹⁰ It reinforced the case for Sealed Mode from a different direction. Where a chatbot conversation predictably contains health or mental-health information, financial information, legal preparation, intimate emotional disclosure or other high-sensitivity material, ordinary terms of service cannot on their own carry the weight of the trust the user places in the system. Policy commitments remain essential, but in these domains they are not, by

⁶ Anthropic, [How people use Claude for support, advice, and companionship](#) (27 June 2025), reporting that affective conversations comprised 2.9% of Claude.ai interactions, with companionship and role-play together under 0.5%, across an analysis of 4.5 million conversations.

⁷ A [Washington Post](#) analysis of tens of thousands of ChatGPT conversations exposed through the product's sharing feature (2025) found that users routinely disclosed details they would not enter into a conventional search engine.

⁸ On reduced fear of judgement and the willingness to confide in a conversational agent, see [Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and its Impact on Emotional Well-Being](#), *Interacting with Computers*, vol. 36 (2024) (finding that users perceived less fear of judgement when disclosing to a chatbot than to a human, even where the intimacy of disclosure did not differ), and the earlier literature reviewed therein on the preference for disclosing sensitive information to a computer rather than to another person.

⁹ Theodore Christakis, "[You Trust Your Chatbot With Everything. Should You? Part 1: How the Controller Uses Your Chat Data](#)," *AI Regulation Papers*, 26-03-2 (March 2026). For an analytical companion piece, see Theodore Christakis, "[New Study Maps the Privacy Gap in Consumer AI and Proposes a Fix](#)", IAPP (March 2026).

¹⁰ Theodore Christakis, "[The Health AI Agent Rush: Five Companies, Your Health Data, and the Governance Questions Nobody is Asking](#)", *AI Regulation Papers*, 26-03-3, (March 2026)

themselves, enough: privacy by design does work that policy alone cannot, and only the two together produce the level of assurance these domains require.

Where Part 1 examined how the provider itself, as data controller, uses chat data, Part 2 takes the next step and asks what happens once the provider is no longer the only party with access. Once a conversation has been generated and stored, it can leave the user's confidential space along **four further pathways**, each the subject of one chapter of this study:

- the provider may refer the conversation to law enforcement on its own initiative;
- the state may compel disclosure through subpoenas, court orders, search warrants, preservation requests or intelligence directives;
- a party to civil litigation may obtain it through discovery, sometimes by the million and sometimes belonging to people who are complete strangers to the case;
- and the conversation may be exposed through a security failure, whether technical, organisational or adversarial.

These four are not an exhaustive catalogue of every way a conversation might reach a third party; they are the four principal routes out of the user's control that this study examines, the ones now generating concrete cases, public controversy and emerging doctrine.

Each pathway is governed by its own legal framework, raises its own governance questions and produces its own kind of harm. But they rest on a single shared fact, and that fact is the connecting thread of this study. A conversation becomes exposed to all four routes for one underlying reason: the provider keeps it. How much damage each route can then do, once it is open, depends on further choices, on whether the conversation is also linked to an account, logged, indexed, personalised, connected to other services or retained in a form that can later be searched; and those choices are not only legal and contractual but managerial and operational, made by product, engineering and security teams and by senior management as much as by lawyers. The same decisions that make a chatbot useful and personalised therefore also make the resulting record preservable, searchable, discoverable, disclosable and exploitable. That convergence, the fact that one retained conversation is at the same time a candidate for referral, a target for compelled access, discoverable evidence and an asset in a breach, is the central object of this study, and to my knowledge it has not been examined as a single phenomenon before.

Valuable work exists on the individual pieces of this picture: on reverse-prompt warrants, geofence and reverse-keyword searches, the third-party doctrine, an AI privilege, generative-AI discovery and the discoverability of AI data. What this study adds is the connection between them.

The pace of these developments is part of what makes them hard to govern. Most of the episodes analysed here occurred within the past year, and several of the most important, the New York Times and OpenAI production orders, the *Heppner* and *Warner* privilege decisions, the *Chatrie* argument, the Tumbler Ridge litigation, the Strasbourg referral, and the recent security and tracking disclosures, surfaced only in the past few months. The doctrines are being applied to these facts, and providers, courts, regulators and users are still forming their expectations, while the practices are already operating at very large scale. This study is therefore best read as a first map of a fast-moving field: provisional in some of its detail, but pressing, because what it describes is already in operation.

The most recent illustration of that pace, registered as this study went to press, is Meta's announcement on 13 May 2026 of *Incognito Chat with Meta AI*. Users of WhatsApp can now chat with Meta AI in a mode in which the provider cannot read the conversation, the closest functional equivalent to the way Meta cannot read WhatsApp messages between two users. The protection is architectural rather than contractual: the chat is processed inside a Trusted Execution Environment in which the provider has renounced, by hardware design, the ability to see the content. This is the first mass-market deployment, in a consumer chatbot, of an architecture in which confidentiality is enforced by design rather than by policy alone, and it bears on several of the analytical frames developed in the chapters that follow. It is examined in a dedicated paper with Peter Swire,¹¹ and its broader significance for this study is taken up in the conclusions.

¹¹ Theodore Christakis and Peter Swire, "[The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI](#)", AI-Regulation Papers 2026-05-4, May 2026. See also WhatsApp Blog, [Introducing Incognito Chat with Meta AI: A](#)

The methodology of Part 2 follows Part 1. The same five providers, ChatGPT, Claude, Gemini, Grok and DeepSeek, are assessed on their public policies, transparency reports and observed behaviour, and the analysis is supported by comparative tables that let the reader see at a glance where each provider stands on the dimensions examined, and where the public record is too thin to support a comparative judgement. Part 2 necessarily reaches beyond the providers, however, to the surrounding legal framework: the constitutional, statutory and procedural rules that determine when, how and on what conditions a chatbot conversation can leave the confidential space the user expects. That analysis is grounded mainly in US law and US transparency reporting, both because four of the five providers are US-based and because the most developed public record of compelled access, the most important discovery battles and the most active doctrinal developments are, for now, in the United States. References to European, Canadian and other law appear where the relevant cases or frameworks have produced concrete material; a complete account of how each non-US jurisdiction governs each pathway is beyond the scope of any single study.

A word on standpoint and limits is in order. This study is written by a European lawyer, and its treatment of US law and procedure should be read as the perspective of an informed external observer rather than the work of a US-licensed practitioner. It does not claim the last word on US discovery, FISA Section 702, privilege doctrine or national-security law, nor on the substantial work already under way in US courts, bar associations, judicial conferences, and the scholarly and practitioner communities. The same limitation applies to Chinese law, EU Member State criminal procedure, UK national-security law and other non-US frameworks, each of which deserves dedicated expert treatment. The aim is to identify and structure a set of connected confidentiality risks that have not, to my knowledge, been examined together.

The study proceeds in four chapters and a set of consolidated conclusions and recommendations.

Chapter 1, “Proactive Corporate Disclosure of Consumer Chatbot Logs”, examines what happens when providers themselves decide, on their own initiative, to refer a user’s conversation to law enforcement. It analyses the three episodes that have defined the field over the past year: the Tumbler Ridge tragedy in Canada; the FSU shooting in Florida; and the Strasbourg case in France, in which a citizen’s chatbot conversation reached the authorities through a chain whose precise routing remains disputed. It then considers the analogy to the *Tarasoff* duty in US clinical practice; the contrast between the standardised CSAM-reporting framework and the unstandardised emergency-referral framework for violent threats; the tension between cryptographic hardening and content-based safety monitoring; and the accountability gap that distinguishes the chatbot setting from the social-media moderation to which it is sometimes compared.

Chapter 2, “When the State Compels Disclosure: Law Enforcement, Intelligence Agencies, and the Surveillance Multiplier”, examines the legal architecture of compelled access. It maps the basic ladder of subpoenas, court orders and search warrants; the procedural safeguards of user notice, resistance and cross-border process; the practical defeat of ephemerality through preservation requests; the emergence of reverse-prompt warrants, in which the state searches not for a known suspect but for whoever entered a given prompt; the foreign-intelligence dimension under FISA Section 702 and the safeguards added by Executive Order 14086; the surveillance significance of memory and agentic AI; the question of separating consumer-facing operations from national-security work, raised by the Anthropic-Pentagon dispute of 2025-2026; and the indispensable but incomplete state of corporate transparency reporting.

Chapter 3, “Discovery: When Chatbot Conversations Become Evidence”, examines the third pathway. It reconstructs the *New York Times v. OpenAI* discovery battle from preservation through production to the March 2026 expansion to eighty-eight million conversations; analyses the failure of OpenAI’s privacy defences and the absence of an “AI privilege” in current US law; examines the *Heppner* and *Warner* decisions of February 2026, which divided early on whether using a consumer AI tool defeats the work-product protection a litigant’s own preparation would otherwise carry; considers why de-identification, protective orders and judicial safeguards reduce but do not remove the confidentiality risk; surveys other pending and emerging cases, including the everyday litigant whose own conversations are sought in an ordinary divorce, employment or business dispute; and asks whether a

[completely private way to chat with AI](#), 13 May 2026; Meta Newsroom, [Introducing a Completely Private Way to Chat With AI](#), 13 May 2026; WhatsApp Help Center, [How Incognito Chat with Meta AI works](#).

“New York Times-style” order could be obtained in Europe, through the disclosure regime of England and Wales, the IP Enforcement Directive, national procedural mechanisms or regulatory investigations.

Chapter 4, “Cyberthreats, Insider Risks and Other Data Breaches”, examines the fourth pathway. It explains why stored chatbot conversations are a distinctive breach target and reviews the public record of warning incidents to date: infrastructure vulnerabilities and software bugs; sharing-design failures and accidental publication; client-side and supply-chain exfiltration; tracker and telemetry leakage from providers’ own platforms; and credential compromise and enterprise leakage. It addresses the new exposure introduced by agentic AI, the insider and vendor dimension, and the legal lens of breach notification, and explains why breach-notification rules, though necessary, are not sufficient for the category of data at stake.

The study closes with **consolidated conclusions and recommendations** addressed to providers, regulators, and oversight bodies on both sides of the Atlantic. Throughout, the analysis is offered in the spirit of opening a debate rather than closing it, and as a contribution to the work that practitioners, scholars and institutions on both sides of the Atlantic have already begun.

1. Proactive Corporate Disclosure of Consumer Chatbot Logs

The people who have historically received intimate, sometimes alarming speech (the doctor, the therapist, the lawyer, the priest, the family member, the close friend) have done so against a background of privilege, professional training, a continuing relationship, or some combination of the three. Hundreds of millions of people now use consumer chatbots, and a great many confide in them in much the same register, but the chatbot has none of those backgrounds to fall back on. When something alarming surfaces in a chatbot conversation, the decisions about what to do (whether to refuse the prompt, suspend the account, route the user to crisis services, alert local police, alert federal authorities, alert authorities in another country, or do nothing at all) are made today, almost everywhere, by the company that operates the chatbot. The criteria are largely not in the public record. The volumes are not published. The error rates are not measured externally.

The legal question this raises is not, in itself, new. Provider-initiated emergency disclosure to the authorities has been part of the legal architecture for electronic communications providers for decades, and the mature processes that providers use to assess claimed emergencies and filter false positives carry over to the chatbot setting rather than starting from scratch. What is distinctive is not any single feature taken alone, since intimacy, ambiguity and scale are each familiar from email, search and social media, but the combination: speech of unusual intimacy, an interpretive ambiguity that resists confident classification, very large scale, and an assessment made by automated systems and content-moderation teams rather than by anyone with the privilege, training or continuing relationship that the traditional confidant brought to the task. That combination is what makes these decisions especially difficult in the chatbot context.

The three episodes that have surfaced over the past year, examined in detail in Section 1.2 (a mass shooting in British Columbia flagged internally and leading to an account ban eight months before the shooting, without referral; a criminal investigation opened by an American state attorney general against the chatbot provider itself for the content of the model's outputs; an alleged cross-border alert chain that put an armed police intervention unit at a Strasbourg apartment within hours of a flagged conversation) are in the public record not because such decisions are rare, since providers probably make them continually and almost always invisibly, but because in these three the consequences became impossible to hide. This chapter examines what those decisions involve, what makes them especially difficult in the chatbot context, and what could be done to bring them under a transparency-and-accountability architecture better suited to what consumer chatbots have already become.

1.1. The legal landscape: existing obligations and what is distinctive about the chatbot context

No jurisdiction currently imposes a general obligation on chatbot providers to monitor user conversations for criminal content or to proactively notify law enforcement of suspicious exchanges.¹² But the legal space is not empty. The relevant frameworks are best understood through the distinction,

¹² This holds for general monitoring duties and for duties to notify law enforcement. A distinct and fast-moving strand of US state legislation does, however, impose narrow obligations on operators of “companion” chatbots to detect and respond to a specific category of content, self-harm and suicidal ideation. California's SB 243 and New York's S-3008C require operators to maintain safety protocols against self-harm content and to direct users to crisis resources, with New York conditioning lawful operation on “reasonable efforts” to detect and address self-harm. These are duties of care owed to the user rather than duties to monitor for criminal content or to notify law enforcement, but they require some degree of content monitoring and intervention, and so sharpen the tension between safety monitoring and confidentiality examined later in this chapter. For an overview, see Justine Gluck, [“Understanding the New Wave of Chatbot Legislation: California SB 243 and Beyond”](#), Future of Privacy Forum (4 November 2025).

classical in this area, between a duty to surveil (which does not exist) and a duty to notify authorities once a provider has become aware of specific facts (which exists in defined circumstances). The remainder of this section sets out the legal frameworks that bear on the duty-to-notify side, examines the *Tarasoff* doctrine that the public debate has invoked alongside them, identifies what is structurally continuous between the chatbot context and earlier emergency-disclosure regimes for electronic communications providers, and identifies what is genuinely new about the chatbot context, which is the analytical focus of the rest of this chapter.

The European Union: Article 18 of the Digital Services Act

Article 18 of the Digital Services Act requires providers of hosting services, when they “become aware” of “any information giving rise to a suspicion that a criminal offence involving a threat to the life or safety of a person or persons has taken place, is taking place or is likely to take place”, to promptly inform the law-enforcement or judicial authorities of the Member State or Member States concerned and to provide “all relevant information available”.¹³ Where the provider cannot identify the Member State concerned with reasonable certainty, Article 18(2) requires notification to the authorities of the Member State in which the provider is established, or where its legal representative resides or is established, or to Europol, or both. Whether AI chatbot providers qualify, in whole or in part, as “hosting services” under the DSA remains unsettled: the DSA was drafted around intermediary services, not general-purpose AI systems as such. Where a chatbot provider stores information provided by users at their request (saved conversations, uploaded files, memory functions, custom GPTs, shared user-created chatbot configurations), the Article 3(g)(iii) definition may become relevant.¹⁴ Article 18 is structured as a triggered notification obligation: it applies where the provider has “become aware” of information giving rise to the relevant suspicion. Article 8 DSA preserves, in parallel, the prohibition on imposing general monitoring or active fact-finding obligations on providers of intermediary services.¹⁵ The legal difficulty lies not only in whether a chatbot provider qualifies as a hosting service under the DSA, but also in identifying when, and through which internal processes, the provider can be said to have acquired the requisite awareness.

The United States: emergency disclosure and CSAM reporting

The US framework combines mandatory and permissive elements that operate on different triggers and warrant separate treatment.

The mandatory element is the obligation, under 18 U.S.C. § 2258A, that providers of “electronic communication service” or “remote computing service” report apparent child sexual abuse material (CSAM) to the National Center for Missing and Exploited Children (NCMEC) when they “obtain actual knowledge” of such content.¹⁶ This is a post-awareness reporting obligation triggered by the existence

¹³ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services ([Digital Services Act](#)), Article 18(1).

¹⁴ On the possible qualification of AI chatbot providers as “hosting services” under the DSA, see DSA Article 3(g)(iii), defining a hosting service as one consisting of the storage of information provided by, and at the request of, a recipient of the service. See also Recital 14 DSA, which clarifies that intermediary services can include services that enable the storage of information, such as cloud computing and web hosting, and that the question is whether the service qualifies functionally, regardless of how it is labelled. The application of this definition to chatbot services remains legally unsettled. Ordinary generation of conversational outputs is not necessarily a hosting activity. The argument is stronger for functionalities involving the storage of user-provided information at the user’s request, such as saved chats, uploaded files, persistent memory, custom GPTs, shared GPTs or other user-created configurations. The argument is weaker for purely provider-side logging retained for safety, debugging, abuse monitoring, improvement, or legal purposes, since such storage may not be “at the request” of the recipient in the sense required by Article 3(g)(iii). The point has not yet, to my knowledge, been adjudicated in relation to general-purpose AI chatbots.

¹⁵ Article 18 DSA should be read together with Article 8 DSA, entitled “No general monitoring or active fact-finding obligations”, which provides that “[n]o general obligation to monitor the information which providers of intermediary services transmit or store, nor actively to seek facts or circumstances indicating illegal activity shall be imposed on those providers.” The Article 18 duty is therefore best understood as an ex post notification obligation triggered by awareness, not as an ex ante obligation to monitor user conversations or proactively detect criminality. This does not preclude providers from becoming aware through voluntary own-initiative measures, user reports, human review, safety escalations, automated moderation pipelines, or other internal processes; nor does it prevent providers from carrying out voluntary own-initiative investigations in accordance with Article 7 DSA. But Article 18, standing alone, does not impose a general surveillance duty.

¹⁶ 18 U.S.C. § 2258A.

of the relevant material, not by a predictive judgment about future harm; CSAM reporting is therefore structurally distinct from threat-based emergency disclosure.

The permissive element is set out in 18 U.S.C. § 2702(b)(8), which addresses the contents of communications, and § 2702(c)(4), which addresses non-content records. Together these provisions allow, but do not require, a provider of electronic communication service or remote computing service to disclose information to a governmental entity if the provider, in good faith, believes that an emergency involving danger of death or serious physical injury to any person requires the disclosure. Several major providers' published policies operationalise these statutory permissions in voluntary commitments to disclose information they believe relevant to an emergency. As under EU law, the framework distinguishes between a duty to notify, which exists in defined circumstances once awareness is established, and a duty to surveil, which does not.

National criminal codes

Several national criminal codes go further than the EU and US frameworks. In France, for example, Article 434-1 of the Penal Code punishes failure to inform judicial or administrative authorities of a crime whose effects can still be prevented or limited; Article 434-3 imposes a more specific reporting obligation concerning deprivation, ill-treatment, sexual assault or other abuse inflicted on minors or vulnerable persons.¹⁷ These provisions are not chatbot-specific, and their application to corporate providers raises questions of knowledge, attribution and statutory exception that fall beyond the scope of this analysis. They illustrate, however, the same feature seen in the EU and US frameworks: the duty is triggered by knowledge of relevant facts, not by a general obligation to seek out potential criminal activity.

The gap

What does not yet exist, in any jurisdiction, is a legal framework drafted with the chatbot context specifically in mind. That is unremarkable in itself, and true of many online services: the statutory definitions of electronic communication service, remote computing service and hosting service are deliberately broad and were intended to reach services not contemplated at the time of drafting. These frameworks therefore reach chatbot providers by their terms, not by analogy. What this chapter examines is accordingly not a gap in legal coverage but a question of calibration: how a body of law developed for messaging, hosting and conventional electronic communications operates when applied to the distinctive combination of features the chatbot context presents.

Provider practices

All five providers examined in this study reserve, in their published policies, the right to disclose user data to government authorities without prior compulsory process where they form a good-faith belief that an emergency involving danger of death or serious physical injury requires the disclosure. Several maintain dedicated reporting pipelines for CSAM and closely related child-safety harms, supported by infrastructure and trained personnel that allow the volume of NCMEC referrals to be processed at scale.

¹⁷ Code pénal, Articles 434-1 and 434-3. Article 434-1 applies to knowledge of a crime that is being committed or could still be prevented. Article 434-3 addresses the specific obligation to report abuse, neglect, or violence against minors and other vulnerable persons.

OpenAI,¹⁸ Anthropic,¹⁹ Google,²⁰ xAI²¹ and DeepSeek²² all publish policy text contemplating proactive disclosures of these kinds, although the legal frameworks under which they operate, and the assurances those frameworks provide, are not equivalent across all five (see the next sub-section on DeepSeek and China).

Public debate often conflates content moderation, account enforcement, emergency disclosure and criminal referral, but they are not the same thing. Account suspension or refusal to answer a particular prompt does not by itself imply referral to law enforcement; it is an internal enforcement decision that may or may not subsequently lead to disclosure. Conversely, the public policies of several providers make clear that a provider may move from internal enforcement to external disclosure where it concludes that an emergency or other reporting threshold has been crossed. The analytical question this chapter engages is how providers define the threshold that activates external disclosure, how much transparency they offer regarding the decision, and what accountability mechanisms govern the process. A comparative overview of the five providers' publicly-stated practices, drawing on the analysis developed throughout this chapter, is presented in Table 1 below.

A note on DeepSeek and China

The analysis of proactive disclosure in this section focuses on US-based providers disclosing to US or European law-enforcement authorities under frameworks that, whatever their gaps, share a broadly common legal grammar: emergency-disclosure thresholds, judicial oversight of compelled access, transparency reporting, and at least a nominal commitment to challenging overbroad demands.

DeepSeek operates under a fundamentally different framework. User data is stored on servers in the People's Republic of China and is subject to Chinese law, including the 2017 National Intelligence Law, whose Article 7 provides that "all organisations and citizens shall, in accordance with the law, support, co-operate with, and collaborate in national intelligence work". The 2021 Data Security Law and the 2021 Personal Information Protection Law further empower Chinese authorities to require data disclosure for national security purposes. I have not located any publicly documented mechanism by which DeepSeek challenges, narrows, reports, or notifies users about such requests. Proactive disclosure by DeepSeek to Chinese authorities therefore operates outside the governance assumptions that underpin this section's analysis: there is no public emergency-disclosure threshold, no known judicial review of access demands, no transparency reporting, and no indication that users are notified under any circumstances. This does not imply that DeepSeek routinely shares consumer chatbot logs with Chinese authorities. It means that the legal architecture does not contain the constraints that, however imperfectly, limit proactive and reactive disclosure by US-based providers. For users outside China who use DeepSeek online, the practical consequence is that their conversations are stored in a jurisdiction where the legal protections analysed throughout this study do not apply.

¹⁸ OpenAI, [OpenAI Government User Data Request Policy](#) (effective 1 January 2026), providing for emergency disclosures where OpenAI has a good-faith belief that disclosure is necessary to prevent danger of death or serious physical injury.

¹⁹ Anthropic, ["What is Anthropic's policy for handling governmental requests for user information?"](#), Anthropic Privacy Center, stating: "We will make an exception to this policy if we believe there is an emergency that may result in imminent physical harm or death, and providing such information without delay may avert that emergency". See also Anthropic, [Government Requests Report](#) (January-June 2025).

²⁰ Google, ["How Google handles government requests for user information"](#), explaining that Google may provide information to a government agency if it reasonably believes it can prevent someone from dying or suffering serious physical harm.

²¹ xAI, [Privacy Policy](#) (effective 10 July 2025), stating that xAI may disclose personal information "in an emergency, to protect the personal safety of xAI, its customers, or any person" and to "comply with laws or to respond to lawful requests and legal process". See also xAI, [Acceptable Use Policy](#) (2 January 2025), stating: "We report suspected child sexual abuse material to the National Center for Missing and Exploited Children".

²² DeepSeek, [Privacy Policy](#) (last updated 10 February 2026), providing that DeepSeek may access, preserve, and share personal data with law-enforcement agencies or emergency services in urgent situations to protect health or life.

The Tarasoff doctrine: conceptual reach and structural distance

The most intellectually visible reference point in US and Canadian debate on a duty to warn is the *Tarasoff* doctrine. The 1976 California Supreme Court decision in *Tarasoff v. Regents of the University of California* held that mental-health professionals have a duty to protect identifiable third parties from serious threats of violence made by their patients, even if doing so requires breaching the confidentiality of the therapeutic relationship.²³ Justice Tobriner's majority opinion concluded with the formulation that has become the doctrine's signature: "the protective privilege ends where the public peril begins". The decision has since been widely adopted across US states with variations, and parallel conceptual moves have shaped Canadian, UK and other common-law debates about when professional confidentiality must yield to identifiable risk. The doctrine has been expressly invoked in the public debate following the Canadian Tumbler Ridge tragedy (see Section 1.2 below); Peter Swire has written and spoken extensively on the relationship between *Tarasoff* and the duty-to-warn questions that arise for AI providers, observing that the previous law concerned an expert psychiatrist with a duty when one patient, whom they knew well, became sufficiently dangerous, and that the scale for an AI provider, and the lack of nuanced knowledge about the individual as assessed by a medical expert, is very different.²⁴

The *Tarasoff* comparison is conceptually instructive and structurally distant from the chatbot context, and both halves of that observation matter. It is conceptually instructive because it foregrounds the analytically prior question that runs through the rest of this chapter: when, and on what terms, may or must the confidentiality of a communication yield to the protection of an identifiable third party? *Tarasoff*'s answer belongs to the same broader legal family as later emergency-disclosure regimes: the idea that confidentiality may yield where a sufficiently serious and identifiable risk to life or safety is present.

It is structurally distant in the sense that the *Tarasoff* doctrine grew up in a clinical relationship with three features that the chatbot context does not share. The *Tarasoff* therapist had a continuing professional relationship with a known patient; the chatbot provider has none of these. The *Tarasoff* therapist was a trained clinician whose professional judgment on questions of dangerousness was the basis for the duty; the chatbot provider's "assessments" are performed by automated classifiers and content-moderation teams operating at scale, not by professionals exercising clinical judgment over time. The *Tarasoff* doctrine operated within the framework of psychotherapist-patient privilege, a doctrinally specific construct that has no counterpart in the legal regime governing chatbot communications. The framework that actually governs the operational regime within which providers decide whether to disclose is therefore not *Tarasoff* but the long-standing emergency-disclosure framework for electronic communication service providers under § 2702 and its EU and national equivalents, in which providers without any professional relationship to their users assess emergency claims on the basis of textual signals and produce disclosures or referrals through internal review processes.

Read this way, *Tarasoff* is useful in the chatbot setting chiefly as a contrast. It names, in a familiar form, the move that emergency-disclosure law also makes: from a default of confidentiality to a duty, or more

²³*Tarasoff v. Regents of the University of California*, 17 Cal. 3d 425, 551 P.2d 334 (Cal. 1976). The facts of the case are instructive. In 1969, Prosenjit Poddar, a graduate student at the University of California, Berkeley, confided to his psychologist, Dr. Lawrence Moore, at the university's Cowell Memorial Hospital, that he intended to kill Tatiana Tarasoff, a young woman with whom he had become infatuated and who had rejected his advances. Dr. Moore took the threat seriously: he notified the campus police and recommended that Poddar be involuntarily committed for psychiatric evaluation. The campus police briefly detained Poddar but released him when he appeared rational. Dr. Moore's supervisor, Dr. Harvey Powelson, then directed that no further action be taken to detain Poddar and ordered that Moore's letter to the police and his therapy notes be destroyed. No one warned Tatiana Tarasoff or her family of the danger. On 27 October 1969, Poddar killed Tarasoff at her home. Her parents sued the university, the therapists, and the campus police. The California Supreme Court, in a landmark 1976 decision, held that mental health professionals have a duty to protect identifiable third parties from serious threats of violence made by their patients, even if doing so requires breaching the confidentiality of the therapeutic relationship. Justice Tobriner's majority opinion concluded: "The protective privilege ends where the public peril begins". The decision has since been widely adopted across US states, with variations: some impose a mandatory duty to warn, others a permissive one, and a few have no statutory position. The *Tarasoff* principle has also influenced broader common-law debates about the circumstances in which confidentiality may yield to the protection of identifiable persons or the public, though different jurisdictions have adopted different formulations. In Canada, the issue is often discussed through the public-safety exception to confidentiality, including *Smith v. Jones*; in the United Kingdom, through the public-interest limits of professional confidentiality. For a comprehensive clinical and legal analysis, see "[Duty to Warn](#)", StatPearls (NCBI Bookshelf).

²⁴Peter Swire's intervention in "[Whispers to Machines: When Chats Become Evidence](#)", Privacy Symposium Panel, 21 April 2026, Venice. See also the discussion of the Tumbler Ridge tragedy in Section 1.2 below.

often in the technology context a permission, to disclose where an identifiable risk to a third party crosses a defined threshold. But it does not supply the operative rule. The regime that both grounds and governs provider-initiated disclosure in the USA is the § 2702-type emergency-disclosure framework, a congressional enactment that carries the policy choice and the legal standard together; *Tarasoff*, with its clinical relationship and its trained professional, is too distant to do more than illuminate by comparison. The chapter's contribution lies in examining what happens when that statutory framework is applied to chatbot conversations, whose combination of intimate register, interpretive ambiguity and very large scale it was not originally calibrated for.

What is continuous, and what is distinctive

Provider-initiated emergency disclosure is not new as a general matter. The § 2702 framework has operated for decades across email, messaging, social-media and cloud-storage providers; the equivalent voluntary practices have a longer history still. Providers have, over that period, developed mature processes for receiving emergency disclosure requests from law enforcement, for assessing the credibility of claimed emergencies, for documenting the basis on which voluntary disclosures are made, and for filtering false positives before referral. The chatbot context inherits this established framework rather than displacing it.

What is genuinely new about the chatbot context, and what gives this chapter its analytical focus, is a combination of features that the established framework was not calibrated for.

The first is the register of the user's interaction with the system. A large and growing number of users confide in consumer chatbots in registers that many of them would not adopt in writing to a human counterpart: as a confidant, a sounding board, an assistant for processing personal difficulty, an interlocutor for working through emotionally charged situations, sometimes as a substitute for human contact they cannot or do not wish to seek elsewhere. The "buddy" framing is not merely a marketing slogan; independent usage research now repeatedly finds personal, advice-seeking and companionship interactions among the common categories of consumer chatbot use,²⁵ and the interfaces are designed to invite exactly that. Conversations that take place in this register can carry an unusually high concentration of sensitive personal disclosure, and the case for treating that material with care in any disclosure decision is correspondingly strong.

The second is interpretive ambiguity. Conversational material with a chatbot blends genuine intent, exploratory thought, fiction, role-play, provocation, system-testing and venting in ways that resist confident classification. A user who writes to a chatbot about violent imagery may be processing emotion, drafting fiction, testing the model's safety responses, expressing genuine hostility, or some combination. The decision whether to refer such a conversation to authorities is harder than the decision to report apparent child sexual abuse material, where it is in the first instance the existence of the material, rather than a prediction about future conduct, that triggers the reporting duty. Even there, as practitioners note, identification and review often demand difficult human judgment; but the trigger differs in kind from a forward-looking assessment of risk. The § 2702 emergency framework requires a good-faith assessment that the relevant emergency exists; the chatbot context is one in which good-faith assessments are unusually difficult to anchor.

The third is scale. Consumer chatbot use is now estimated at over one billion weekly users globally. Mature emergency-disclosure regimes in adjacent contexts have operated at large scale before, and providers have developed processes to filter false positives before they become referrals. The chatbot context nevertheless adds, on top of that established practice, a base of conversational material that is unusually intimate and unusually ambiguous. Even modest error rates applied to a base of that size and that character could produce a meaningful absolute volume of erroneous referrals, and the threshold for external disclosure must be calibrated with that base in mind. The point is not that the scale issue is new in the abstract; it is that the conjunction of unusually intimate material, interpretive ambiguity and unprecedented user volume is specific to this setting.

The fourth is the absence of professional expertise. The assessments through which a chatbot provider becomes "aware" of relevant facts are performed in the first instance by automated classifiers trained on

²⁵ See the introduction to this Part 2 as well as Part 1 of the study.

conversational data, with content-moderation teams reviewing flagged material at later stages of the moderation pipeline. The provider does not know the user, has no clinical relationship with the user, and cannot draw on professional judgment of the kind that animates the duty-to-warn doctrines developed for clinical settings. This is the same point that distinguishes the chatbot referral decision from the clinical duty-to-warn model: the assessments are made by automated systems and content-moderation teams, not by a clinician exercising professional judgment over a continuing relationship. It is not a point of difference from the content-moderation and emergency-disclosure operations of adjacent online services, with which the chatbot referral decision shares much of its method, its signals, and often its personnel. The contrast that matters here is with the clinical model the public debate invokes, not with the practice of other providers.

Two analytical categories of provider-initiated disclosure

The episodes analysed below illustrate two analytically distinct categories of provider-initiated disclosure that this chapter addresses.

The first is the emergency-framed disclosure: a provider, having become aware of conversational content that it assesses in good faith as indicating an imminent threat to life or serious physical injury, contacts the relevant authorities under the § 2702 permissive disclosure provisions or their EU and national equivalents.

The second is the non-emergency proactive referral: a provider, having become aware of conversational content that it assesses as evidence of past or planned criminal conduct without an emergency framing, refers the matter to the relevant authorities. The legal basis for this second category may vary more substantially than for emergency disclosure, especially where contents of communications are involved. It may depend on the nature of the information disclosed, the provider's statutory status, applicable statutory exceptions, and the law of the receiving jurisdiction. The point for present purposes is that both categories raise threshold-setting and accountability questions.

1.2. When moderation becomes referral: three recent episodes

Three recent episodes from 2025-2026 show how these legal categories become concrete when a problematic conversation moves from a moderation concern to an external referral. They do not exhaust the field of provider-initiated disclosure, nor do they purport to settle the legal questions that arise in it; they are useful to this analysis because each presents, with some clarity, a distinct decision point at which a provider's internal moderation pipeline meets the question whether to involve external authorities, and what happens when it does.²⁶

The Tumbler Ridge tragedy: detection without referral (Canada, February 2026)

In February 2026, a mass shooting in the small community of Tumbler Ridge, British Columbia, killed eight people and left others with catastrophic injuries. Among the wounded was 12-year-old Maya Gebala, who was shot in the head and neck while trying to lock the school library door to protect other students, and who suffered catastrophic brain injuries. According to press reports and the subsequent lawsuits, the perpetrator had used ChatGPT extensively in the months before the attack.^{27,28} The Gebala family's complaint alleges that the shooter used ChatGPT to research and plan the attack, and that the content of his conversations was sufficiently alarming that approximately 12 OpenAI employees who reviewed the account identified his posts as "indicating an imminent risk of serious harm to others" and recommended that Canadian law enforcement be informed, but that these recommendations were

²⁶ On the distinction between emergency-framed disclosure under § 2702 and the broader category of non-emergency proactive referrals, and on the legal authorities that bear on each, see Section 1.1 above.

²⁷ "Altman apologizes after OpenAI failed to alert police before Tumbler Ridge killings", AP, 25 April 2026.

²⁸ See Global News, "OpenAI's handling of Tumbler Ridge shooter info opens regulation questions", 24 February 2026; Global News, "AI regulation and Canadians' privacy in wake of Tumbler Ridge shooting", 13 March 2026.

“escalated to leadership” and “rebuffed”.²⁹ As noted below in Section 1.6, lawsuits by several other victims’ families followed.

The case has become the first major public test, of a chatbot provider’s referral decision in a violent-incident context, and the episode has surfaced the analytical question whether the threshold OpenAI applied at the relevant time satisfied the company’s own emergency-disclosure standard or, more broadly, the standards that ought to govern such decisions. In its 26 February 2026 letter to Canadian ministers, OpenAI publicly acknowledged that in June 2025 it had shut down a ChatGPT account belonging to the later perpetrator after its automated systems detected a usage-policy violation (later reported by the press as specifically related to the “furtherance of violent activities”). The account was escalated to human review to determine both whether the policy had been violated and whether the matter warranted referral to law enforcement. At that time, as OpenAI states in that letter, it concluded that it had not identified “credible and imminent planning” sufficient to meet its referral threshold, and therefore did not alert law enforcement. After the February 2026 tragedy, OpenAI stated that under its enhanced law-enforcement referral protocol, the same account would be referred today if discovered now.³⁰

The importance of Tumbler Ridge lies in what it reveals about how provider referral decisions are actually made. The hard part is not detecting troubling language but turning ambiguous material into a judgement about real-world risk, and that judgement is reached through automated classifiers, content-moderation review and escalation to senior staff, not by a clinician who knows the user. The provider’s knowledge is different in kind from a therapist’s: usually thinner, because there is no relationship, but occasionally more explicit, because a user may put in writing to a chatbot operational detail, how to obtain or load a weapon for instance, that a patient would rarely raise with a therapist. Either way, the decision-maker and the basis for the decision are not the ones the clinical duty-to-warn doctrines were built around, which is the point developed in Section 1.1.

The case also shows that internal referral thresholds are not fixed. After Tumbler Ridge, OpenAI stated that under its enhanced protocol the same account would be referred today, a non-trivial shift in its stated practice. What standard, if any, should govern these thresholds going forward, and whether it should be developed within the existing emergency-disclosure frameworks (which already apply to AI providers as electronic communication or remote computing services), through new chatbot-specific obligations, or through some combination, are questions to which the rest of this chapter returns.

The FSU shooting: post-incident cooperation and its sequel (Florida, April 2025-April 2026)

The Florida State University campus shooting of 17 April 2025, in which two people were killed and six others wounded, began as an ordinary instance of post-incident cooperation with an active criminal investigation. Producing evidence once an investigation is underway is the most commonplace form of provider engagement with law enforcement, governed by ordinary subpoena and warrant practice rather than by the emergency-disclosure framework discussed in Section 1.1.

What makes the FSU episode relevant here is not the post-incident cooperation, which is routine, but the criminal investigation that followed, and its bearing on disclosure is indirect though important. The investigation shifted the focus from the suspect’s use of ChatGPT to whether the provider itself could face criminal exposure for the content of the model’s outputs. Threats to hold a service criminally exposed for harms connected to its platform are not new, as the Backpage prosecution in the United States and the investigation of Telegram’s Pavel Durov in France show; what is comparatively novel is that this theory targets the model’s own generated advice rather than user-posted content. The reason it matters for a paper about disclosure is the incentive it creates: a provider that fears being blamed for an attack it failed to report has a strong reason to refer earlier and more often, which is exactly the over-referral risk this chapter weighs against the privacy cost of disclosure.

²⁹ *Gebala v. OpenAI*, filed in B.C. Supreme Court, 9 March 2026. See Business in Vancouver, “Tumbler Ridge shooting victim sues OpenAI”, 10 March 2026.

³⁰ Ann M. O’Leary, Vice President of Global Policy, OpenAI, [letter to Canadian ministers \(26 February 2026\)](#), stating that OpenAI banned the later Tumbler Ridge perpetrator’s account in June 2025 after automated detection and human review, but did not then find “credible and imminent planning” sufficient for a referral; further stating that under OpenAI’s enhanced law-enforcement referral protocol the account would be referred today.

According to press reports and court filings, the suspected shooter, Phoenix Ikner, was in extensive communication with ChatGPT before the attack. Court records reportedly list more than 270 OpenAI photos and ChatGPT conversations as exhibits in the criminal case, referenced in the State's answer to discovery dated 31 July 2025.³¹

The timeline requires careful attention. The FSU shooting occurred on 17 April 2025. OpenAI's public statement confirmed that "after learning of the incident in late April 2025", it "identified a ChatGPT account believed to be associated with the suspect, proactively shared this information with law enforcement and cooperated with authorities".³² The word "proactively" in OpenAI's statement refers to the company's initiative in sharing information after learning of the shooting from external sources; it does not mean that OpenAI alerted authorities before the attack. This is therefore post-incident cooperation, not pre-incident referral.

On 21 April 2026, Florida Attorney General James Uthmeier announced a criminal investigation into OpenAI, alleging that ChatGPT "advised" Ikner on weapons, ammunition, "what time of day would be appropriate for the shooting to interact with more people, and where on campus would be the place to encounter a higher population". Uthmeier stated at a press conference that "if ChatGPT were a person, it would be facing charges for murder". The Florida Office of Statewide Prosecution has subpoenaed OpenAI for records of its policies and training materials regarding threats to harm others, cooperation with law enforcement, and reporting of possible past, present, or future crime. OpenAI responded that the shooting "was a tragedy, but ChatGPT is not responsible for this terrible crime", and that "ChatGPT provided factual responses to questions with information that could be found broadly across public sources on the internet, and it did not encourage or promote illegal or harmful activity". Ikner has pleaded not guilty, and his trial is set to begin in October 2026.³³

The episode therefore began as routine post-incident cooperation but has surfaced the liability pressure just described. Whatever becomes of the Florida investigation, including the real possibility that it is dropped or resolved without charges, the fact that a state attorney general opened it at all is part of the incentive environment in which providers now calibrate their referral decisions.

The Strasbourg case: cross-border escalation in practice (France, April 2026)

The third episode shows how cross-border escalation can operate when a provider decides to alert authorities, on facts known so far only through French press reporting. According to French press reports from early April 2026, a 37-year-old man in Strasbourg told ChatGPT that he wanted to acquire a Glock pistol in order to kill an intelligence officer from the CIA, the Mossad, or the DGSI.³⁴ The conversation was reportedly detected by OpenAI's moderation systems. According to multiple French press accounts, OpenAI alerted the FBI, which in turn transmitted the alert to French authorities via the PHAROS platform (managed by the OCLCTIC, France's central office for combating information-technology crime).³⁵ The RAID, France's elite police intervention unit, intervened at the suspect's home in the Koenigshoffen quarter of Strasbourg on the same day (Friday 3 April 2026).³⁶ No weapon was found. The suspect was placed in custody, then released the following day and hospitalised on a compulsory basis due to psychiatric antecedents. The suspect later stated that he had wanted to "test the reliability and surveillance of artificial intelligence" rather than plan any attack. The case was closed without further action.

³¹ See First Coast News, "Family of FSU shooting victim claim attacker received advice from ChatGPT", 7 April 2026; "Alleged Florida State shooter used ChatGPT to plan attack, victim's attorneys claim", WCTV / Gray News, 8 April 2026.

³² OpenAI statement, cited in NewsNation, "FSU shooting: What we know about ChatGPT connection", April 2026.

³³ CNN, "Florida attorney general launches criminal investigation into ChatGPT maker OpenAI after deadly FSU shooting", 21 April 2026; NPR, "Florida AG launches criminal investigation into ChatGPT over FSU shooting", 21 April 2026. The Florida Office of Statewide Prosecution has subpoenaed OpenAI for records including "policies and internal training materials regarding cooperation with law enforcement, user threats of harm to others and user threats of harm to self, as well as lists of employees and an organizational chart": see UPI, "Florida AG opens criminal investigation into OpenAI, ChatGPT over FSU shooting", 21 April 2026.

³⁴ Le Parisien, "Il menace de tuer des agents du renseignement sur ChatGPT : le Raid l'interpelle à Strasbourg après un signalement du FBI", 7 April 2026.

³⁵ It should be noted that OpenAI has not publicly confirmed the case nor this specific chain. Persons familiar with the company's practices have indicated that chatbot providers **generally refer safety concerns to local law-enforcement authorities** in the relevant jurisdiction rather than routing them through US federal agencies.

³⁶ See INFO.FR, "ChatGPT alerte le FBI : un Strasbourgeois déclenche le RAID", 6 April 2026; L'Est Républicain, "Strasbourg : il évoque un meurtre devant une IA, le FBI alerte et le RAID intervient 6 April 2026.

An interesting, but *not confirmed* aspect of this case concerns the cross-border data element. If the French press accounts are accurate,³⁷ the case provides a publicly documented illustration of how a cross-border safety-alert chain operates in the chatbot context: consumer chatbot conversations trigger automated moderation flags, those flags are escalated within the provider, the provider allegedly alerts US federal authorities, and the alert is routed to local law enforcement in the user's country. Cross-border escalation chains are not themselves new. US providers have for decades passed emergency information to foreign authorities, often through the FBI's network of Legal Attachés and the long-established 24/7 channels for urgent cross-border assistance; the financial-intelligence and counter-terrorism contexts work with broadly similar mechanics. What the Strasbourg case might add is an example of the same kind of mechanics in the consumer chatbot setting.

Two cautions follow. First, the referral should not be judged by what the authorities did with it. A provider cannot foresee whether a flag will produce a proportionate response or a heavy-handed one, so the decision to refer has to be assessed on what the provider knew at the time, not on the downstream outcome; the hospitalisation and closure of the file here may well have been appropriate on facts that are not public. Second, and for the same reason, the receiving authority's protocols matter as much as the provider's threshold. How police triage a chatbot flag, check it against other information, and decide whether to speak to the user before sending in an intervention unit is an independent variable in the proportionality of the whole chain. Those protocols are not new in form; authorities have refined them over decades for tips arriving through other channels. What is new is the channel, and the difficulty of reading the material that travels down it, which Sections 1.1 and 1.7 address.

What the three episodes illustrate

The three episodes are useful taken together because each illustrates a different point in the lifecycle of a moderation-to-referral decision. Tumbler Ridge illustrates the decision whether to refer in the absence of an immediate emergency, where the legal framework permits voluntary disclosure but the standards that govern such disclosure are largely set by the provider rather than by external authority. The FSU episode illustrates the most commonplace form of provider engagement (post-incident cooperation), but its Florida sequel illustrates a different and broader question: whether a provider can be held criminally accountable for the substantive content of its model's outputs. The Strasbourg case illustrates the rapidity of intervention following a proactive referral.

Three cases do not prove a structural claim, and private firms are not generally required to publish the criteria behind decisions of this kind; the absence of published criteria is not, by itself, evidence of a regulatory gap. What the three episodes do is make a particular configuration concrete. The decisions, whether to flag, whether to refer, whether to cooperate and on what terms, are taken by automated systems and human reviewers inside a private moderation pipeline; they affect a global user population that cannot see the criteria being applied; the underlying material is intimate, ambiguous and voluminous in the combination Section 1.1 describes; and the consequences for the individual can be severe, up to armed intervention, criminal investigation and compulsory hospitalisation. The claim is not that this amounts to a regulatory failure requiring legislation. It is narrower: at present these decisions cannot be evaluated from outside at all, because the criteria, the volumes and the error rates are not published, and it is that opacity, rather than any proven failure, that the rest of the chapter argues should change.

The Canadian legislative response

The Tumbler Ridge tragedy has triggered a significant legislative and policy debate in Canada. Canada currently has no federal law requiring AI companies to report potentially violent users to law enforcement, and, as the proposals below show, the debate is directed at transparency and oversight rather than at creating such a reporting duty.³⁸ Within days of OpenAI's acknowledgment, the Minister of Artificial Intelligence and Digital Innovation, Evan Solomon, summoned OpenAI officials to Ottawa and stated that "Canadians expect online platforms, including OpenAI, to have robust safety protocols

³⁷ But see note 35 above.

³⁸ CBC News, "[What the Tumbler Ridge mass shooting reveals about regulating AI](#)", 2 March 2026.

and escalation practices in place”.³⁹ The Globe and Mail reported that experts called for the forthcoming federal online harms bill to address AI chatbots specifically and to create a framework for reporting credible threats, including mandated transparency reporting on how often internal flags lead to external referrals and regulatory audit authority to assess whether escalation protocols are robust and consistently applied.⁴⁰ Michael Geist, testifying before the House of Commons Standing Committee on Industry, Science and Technology in March 2026, argued instead for an AI Transparency Act that would require publicly accessible corporate policies, mandated transparency reports on law-enforcement referrals, and disclosure of how flagging systems operate, while cautioning that extending the Online Harms Act to AI chatbots would require “gutting the very privacy protections the government added” to the earlier bill.⁴¹ Academics have drawn a parallel with the Tarasoff principle in Canadian mental-health law, which imposes on therapists a duty to warn when a patient poses a credible threat to an identifiable person, and have asked whether a comparable duty should apply to AI providers, noting that “the people making these assessments are software engineers and content moderators, not forensic psychologists”.⁴² As Section 1.1 sets out, the Tarasoff parallel is conceptually instructive but structurally distant from the chatbot context, in which the relevant assessments are performed by automated classifiers and content-moderation pipelines rather than by individual professionals exercising clinical judgment. A British Columbia inquiry has been announced to examine the role AI may have played in the shooting.⁴³

The Canadian debate is currently illustrating the governance questions that every jurisdiction will eventually face. The range of possible institutional responses is broader than the options most prominently discussed in the Canadian press: it extends beyond a binary choice between corporate self-regulation and, at the far end, a legislative duty not merely to report but to monitor for and detect reportable content, a duty that, as Section 1.1 notes, no jurisdiction currently imposes. Other models include structured post-hoc review of referral decisions, mandated transparency reporting (on the CSAM model), regulatory audit authority, industry-developed standards subject to independent certification, or some combination of these.⁴⁴ This chapter does not prescribe a specific institutional design. It identifies the categories that need to be distinguished, the stakes involved, and the transparency and accountability questions that follow.

1.3. OpenAI after the Canadian tragedy

In its 26 February 2026 letter to Canadian ministers, OpenAI publicly described how it had handled the perpetrator’s account before the Tumbler Ridge tragedy and how it intended to adjust its practices going forward.⁴⁵ The letter has analytical value because it is one of the first publicly available statements by a major chatbot provider to set out, in some operational detail and in response to a specific incident, how an internal referral process worked in a concrete case and how it is being modified.

OpenAI stated that it had made its referral criteria “more flexible”, that mental-health, behavioural and law-enforcement experts were now helping assess difficult cases, that it would create direct points of contact with Canadian law-enforcement authorities, and that it would strengthen systems designed to detect repeat policy violators who attempt to create new accounts after a ban. Three observations follow.

First, the letter does not establish a referral protocol that did not previously exist. Major US chatbot providers have published emergency-disclosure policies for some time (see Section 1.1 above), and providers’ transparency reporting has documented their handling of emergency-disclosure and

³⁹ Canada’s National Observer, “[After Tumbler Ridge, the Carney government needs to lead on AI regulation](#)”, 25 February 2026.

⁴⁰ The Globe and Mail, “[Online harms bill needs framework for reporting threats in AI chats, experts say](#)”, 23 February 2026.

⁴¹ Michael Geist, “[Setting Canada’s AI Policy Priorities: My Appearance Before the Standing Committee on Industry, Science and Technology](#)”, 23 March 2026.

⁴² The Conversation, “[Danger was flagged, but not reported: What the Tumbler Ridge tragedy reveals about Canada’s AI governance vacuum](#)”, 24 February 2026. The Tarasoff parallel is also discussed in The Conversation, “[OpenAI’s safety pledges in the wake of Tumbler Ridge aren’t AI regulation, they’re surveillance](#)”, 7 April 2026, arguing that OpenAI’s enhanced referral protocol is “the same unaccountable architecture with a faster trigger”.

⁴³ Global News, “[AI regulation and Canadians’ privacy in wake of Tumbler Ridge shooting](#)”, 13 March 2026, noting that “an inquiry set to take place in B.C. will also look at the role artificial intelligence may have played in the shooting”.

⁴⁴ Ann M. O’Leary, Vice President of Global Policy, OpenAI, [letter to Canadian ministers \(26 February 2026\)](#).

⁴⁵ Ann M. O’Leary, Vice President of Global Policy, OpenAI, [letter to Canadian ministers \(26 February 2026\)](#).

proactive-referral matters across multiple reporting cycles. What the letter adds to the public record is therefore not the existence of a referral protocol, but a specific account of how that protocol operated in a specific case (the June 2025 ban without referral), and a description of specific operational adjustments made afterwards (the integration of mental-health expertise, direct Canadian-jurisdiction contact, and repeat-violator detection across accounts).

Second, the letter illustrates the dynamics by which internal referral thresholds adjust under public pressure, but it does not, on its own, necessarily demonstrate a broader trend. OpenAI's June 2025 conclusion that no "credible and imminent planning" had been identified was, on the information available at the time, a defensible application of its then-existing emergency-disclosure threshold. Its February 2026 statement that the same account would now be referred is an adjustment to the operational application of that threshold, not a change in the underlying legal framework. Emergency disclosure based on incomplete information has a long history across email, messaging and social-media providers, and adjustments to internal referral practices following high-profile incidents are not in themselves new. How widely the kinds of adjustments described in the OpenAI letter will be made by other providers, whether they will converge on a common chatbot-specific standard, and how durable the adjustments prove over time, are open empirical questions to which one event involving one provider cannot give a confident answer. The point is rather that the letter offers an unusually detailed glimpse into the operational reasoning of a major provider on a chatbot-context referral decision, which is useful for analytical work even where it cannot settle the larger questions.

Third, the letter illustrates how hard the underlying judgement is. Conversational systems capture fragments of intent, fantasy, distress, experimentation, bravado and planning, but rarely a complete evidentiary picture. Converting that into a judgement about real-world risk is difficult in ways largely shared with adjacent emergency-disclosure contexts. What Section 1.1 identifies as particular to the chatbot setting is not any one of these features on its own but their combination: an intimate and ambiguous register, read without the continuing relationship through which a clinician, or even a friend, could tell what a given utterance is doing, and at very large scale. The fuller comparison with prior emergency-disclosure contexts, including social-media moderation and ordinary search, is reserved for Section 1.7.

Beyond the letter itself, OpenAI's broader safety materials, and in particular its Preparedness Framework, help to explain how a chatbot-specific referral process can function in practice. The Framework describes monitoring of harmful user requests, model responses and aggregate usage patterns using auxiliary models trained to detect harmful tasks and jailbreak attempts; blocking unsafe requests where possible; escalating to human review where needed; expanding human monitoring and investigation capacity; and maintaining escalation channels with law enforcement.⁴⁶

One feature of this architecture warrants a short observation here, because it bears directly on the accountability questions that the rest of the chapter examines. The initial triage in the pipeline is itself performed by an AI system: an auxiliary classifier evaluates the input or output of the primary model to determine whether the conversation presents a risk sufficient to escalate to human review.⁴⁷ This AI-on-AI safety architecture, including the absence of publicly available accuracy metrics, false-positive rates, or independent audit, is one of the accountability questions to which the chapter returns in its discussion of transparency reporting (Section 1.7).

Taken as a whole, the letter is a useful but limited data point. It is useful because it brings into the public record a more granular account of one major chatbot provider's internal referral architecture than was previously available, and because it places that account in the context of a concrete decision (the June 2025 non-referral) that the provider has publicly acknowledged would today be decided differently. It is limited because it cannot, on its own, settle either the empirical question of how providers should set their referral thresholds or the legal question of how those thresholds should relate to existing emergency-disclosure frameworks.

⁴⁶ OpenAI, [Preparedness Framework version 2](#) (15 April 2025).

⁴⁷ *Ibid*, at Section 3, describing "usage monitoring" that includes "auxiliary models trained to detect harmful tasks and jailbreak attempts". On the broader question of AI-on-AI safety monitoring, see also Anthropic, [Transparency Hub – System Trust Reporting](#) (29 January 2026), describing the use of "detection classifiers" as part of its CSAM detection pipeline. Neither provider has published accuracy metrics, false-positive rates, or independent audit results for these classifier systems.

1.4. The mounting legal pressure on providers, and its effect on referral incentives

The referral decisions described above are not taken in a vacuum. Consumer chatbot providers now face a fast-growing body of civil litigation, criminal investigation and proposed legislation directed at the harms their systems are alleged to have caused or facilitated. That environment bears on proactive disclosure in a specific way: it shapes the incentives under which a provider decides whether to refer a user to the authorities. The effect is not a simple push towards more reporting. As this section argues, the same liability pressure pulls in both directions at once, and it is that two-sided pressure, rather than any one-way “incentive shift”, that makes the calibration of referral thresholds so difficult.

The litigation and enforcement environment

On the civil side, providers are defending wrongful-death and product-liability suits arising from harm both to others and to users themselves. The most developed are the companion-chatbot suicide cases: the Garcia litigation over the death of fourteen-year-old Sewell Setzer III, in which a federal court allowed negligence and product-liability claims against the developer to proceed past the motion-to-dismiss stage, and comparable actions brought by the families of other minors and, more recently, of adult users.⁴⁸ The Tumbler Ridge families have since sued OpenAI and its chief executive in California federal court, alleging that the company chose not to notify police, and the family of one FSU victim has signalled a civil claim of its own.⁴⁹ These are not the first suits to test a service provider’s responsibility for harms connected to its platform, but their shared theory is what matters here: that the provider knew, or should have known, of a serious risk and failed to detect, warn, escalate or refer accordingly.

On the criminal and regulatory side, state attorneys general have begun to apply criminal and consumer-protection authorities to providers. Florida opened a criminal investigation of OpenAI in connection with the FSU shooting (Section 1.2 above); New Mexico has disclosed in February 2026 an investigation of unnamed chatbots; and California and Arizona have investigated xAI’s Grok over sexually explicit outputs. Federal proposals to treat AI systems as “products” for liability purposes would, if enacted, widen the doctrinal channels open to plaintiffs and regulators alike. Much of this remains at the allegation or investigation stage, and we don’t know whether any particular claim might succeed. The point for this chapter is that providers are now setting their referral thresholds in the shadow of liability arriving from several directions at once.

The documentation paradox

One feature of this environment deserves to be isolated, because it bears directly on how providers build and record their referral processes. A successful liability claim of the kind just described will often turn on knowledge: on showing that the provider knew of a risk and proceeded regardless.⁵⁰ It is important to be precise about which knowledge is at stake, because two quite different things travel under the word. In the product cases, the relevant knowledge is knowledge about the system: that the model can generate a given category of harmful output. In the referral context that is the subject of this chapter, the relevant

⁴⁸ Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla.). In May 2025 the court denied the defendants’ motion to dismiss in large part, allowing the negligence and strict-product-liability claims to proceed: [Tech Justice Law, case summary](#); [docket](#). The Setzer matter, with related actions in Colorado, New York and Texas, was reported settled in January 2026. See also the wrongful-death action over the death of sixteen-year-old Adam Raine, filed against OpenAI and Sam Altman in the Superior Court of California in August 2025, and the suits by adult users and their families filed in California in November 2025.

⁴⁹ On 23 April 2026 Sam Altman published a personal apology for the failure to alert law enforcement to the account banned in June 2025; on 29 April 2026 eight wrongful-death and personal-injury actions were filed against OpenAI and Altman in California federal court on behalf of the Tumbler Ridge victims’ families. Attorneys for the family of Robert Morales, one of the victims of the April 2025 FSU shooting, have indicated that they will bring a civil action against OpenAI. See [“Families of Tumbler Ridge shooting victims sue OpenAI and CEO Sam Altman”](#), CNN, 29 April 2026; [“Altman apologizes after OpenAI failed to alert police before Tumbler Ridge killings”](#), AP, 25 April 2026.

⁵⁰ As practitioners have observed, a criminal case against a consumer AI provider will turn on proof of knowledge that the system could produce the relevant harmful output and that the provider deployed or maintained it nonetheless: see Sarah Kay Wiley, Managing Director, Tech Justice Law quoted in Emma Whitford, “Criminal probe of OpenAI ramps up pressure on chatbot developers in US”, MLex (21 April 2026).

knowledge is narrower and case-specific: what the provider knew, or recorded, about a particular user's conversation when it decided whether to refer. It is the second kind that creates a paradox. The more carefully a provider reviews individual accounts and documents its escalation decisions, the richer the record a future plaintiff or prosecutor can cite to argue that it saw a specific risk and did nothing. This is the familiar safe-harbour dilemma of online-intermediary law in a new setting: the provider that invests most in internal review also generates the most evidence for use against itself.⁵¹ Until the law develops a category that protects documented, good-faith safety effort while preserving liability for the knowing disregard of specific and serious risk, the incentive runs the wrong way, towards thinner records rather than better ones.

A two-way incentive, and why it matters for disclosure

The pressures just described do not point in a single direction.

On one side, the prospect of being blamed after a tragedy, that the provider detected the risk and failed to act, gives it a strong reason to refer early and often, and the more its own monitoring detects, the harder that reproach is to answer.

On the other side, a pipeline that routinely converts ambiguous conversation into law-enforcement attention carries its own costs: it exposes users to the consequences set out in Section 1.2, it burdens innocent users caught by false positives, it strains the trust on which the product depends, and a high-volume referral system is itself a moderation operation whose errors will in time generate their own complaints and claims. **A provider therefore has no safe default. Referring too readily and referring too rarely each carry real and growing legal and reputational risk.**

That is what this environment establishes for the chapter. Because the incentives cut both ways, the governance question cannot be answered by a simple instruction to refer more or to refer less; either, applied across hundreds of millions of users, produces large absolute harms of a different kind. What the situation calls for instead is visibility into how the threshold is actually set and applied. Those are the **transparency and accountability questions** that the rest of this chapter takes up, beginning with the one domain, child-safety reporting, in which a standardised framework already exists.

1.5. Child safety as the most standardised domain of proactive reporting

If provider-initiated referrals based on violent-threat detection remain comparatively opaque, child-safety reporting is, within the category of provider-initiated proactive disclosures⁵², far more standardised. OpenAI's trust and transparency hub reports 107,817 CyberTipline reports to the National Centre for Missing and Exploited Children (NCMEC) and 107,667 pieces of content reported to NCMEC for July–December 2025.⁵³ Anthropic's transparency hub states that on its first-party services it uses hash-matching technology and detection classifiers to detect and report CSAM that users may upload, and reports 5,005 pieces of content reported to NCMEC for July–December 2025.⁵⁴ xAI's Acceptable Use Policy explicitly states that it reports suspected child sexual abuse material to NCMEC.⁵⁵

Google's public child-safety documentation is even more explicit about the underlying logic of detection and reporting. Google explains that it devotes substantial technological and human resources to deterring, detecting, removing, and reporting child sexual exploitation content and behaviour; that it relies on a mixture of automated systems, classifiers, and human review; and that its child-safety transparency reporting includes data on reports made to NCMEC and hashes shared to help identify abusive material at scale.⁵⁶ Although this documentation is Google-wide rather than Gemini-specific, it

⁵¹ The dynamic is familiar from the literature on intermediary liability and statutory safe harbours, in which the "knowledge" that exposes a provider to liability can be generated by the very monitoring undertaken to prevent harm.

⁵² Reporting on responses to government demands such as subpoenas, warrants, and FISA process is also highly standardised, in some cases by statute, but operates on a different axis and is the subject of Chapter 2.

⁵³ OpenAI, [Trust & transparency hub](#) and [Report on Child Safety Reporting \(July–December 2025\)](#), reporting 107,817 CyberTipline reports and 107,667 pieces of content reported to NCMEC.

⁵⁴ Anthropic, [Transparency Hub – System Trust Reporting](#) (29 January 2026).

⁵⁵ xAI, [Acceptable Use Policy](#) (2 January 2025).

⁵⁶ Google, ["Fighting child sexual abuse online"](#).

remains highly relevant because Gemini is embedded in the broader Google legal and safety architecture.

The public record is therefore strongest for CSAM and closely related child-exploitation scenarios. By contrast, for terrorism, extremism, suicide, or generalised violent ideation, providers often describe refusal systems, safety protocols, or emergency-disclosure rights without publishing comparably granular statistics on proactive law-enforcement referrals.

That asymmetry is worth noting, though it should be understood in context. Much CSAM detection benefits from characteristics that violent-threat assessment often lacks: known images and videos can be hash-matched against established databases and reporting infrastructures; many core categories of material are more readily classifiable than ambiguous threat language; and the legal obligation to report is clear and well-established. The contrast should not, however, be overstated. Even within CSAM, classification at the borderline is more contested than the headline reporting figures might suggest. In the United States, courts continue to rely on the six-factor framework developed in *United States v. Dost* to assess whether material falls within the “lascivious exhibition” element of the federal child-pornography statute, an inherently fact-intensive inquiry that considers, among other things, the focal point of the depiction, the setting, the pose, the degree of clothing or nudity, and whether the image appears intended to elicit a sexual response.⁵⁷ The factors are not exhaustive and no single factor is dispositive. The issue has become still more practically significant as synthetic and AI-generated material increases the volume of novel or non-hash-matchable content and pushes more cases toward classifier-based and human judgement rather than simple matching against known hashes.⁵⁸ CSAM, in other words, has its own zones of classification ambiguity, even if its central core remains more determinate than the violent-threat domain.

If classification at the borderline is contested even for CSAM, it is more contested still for violent-threat assessment and adjacent categories such as terrorism, extremism, or self-harm referrals. These assessments depend almost entirely on interpreting ambiguous conversational material, in which the level of certainty, the imminence of the harm, and the appropriate response are all contested. Nor do these domains benefit from an equivalent institutional architecture. There are, of course, legal definitions of terrorism, threats, incitement and related offences, and there are specialised reporting and coordination mechanisms in some jurisdictions. But there is no single statutory reporting framework equivalent to § 2258A, no NCMEC-style clearinghouse for chatbot threat referrals, no universally accepted classifier benchmark, no centralised hash-matching infrastructure comparable to known-CSAM matching, and no agreed multi-factor judicial test capable of translating ambiguous chatbot conversations into referral decisions. Comparing transparency standards across these domains therefore requires caution. Nevertheless, the basic principle remains sound: a governance mechanism capable of triggering law-enforcement referrals should be accompanied by some form of aggregate reporting. The question is what form such reporting should take, given the far greater complexity of the violent-threat and self-harm domains.

For CSAM, the figures reported by OpenAI and Anthropic above allow at least a baseline assessment of the scale and trajectory of provider-initiated reporting in this category. For violent-threat referrals, no equivalent data exists⁵⁹ (Table 1 below, in the closing portion of this chapter, summarises this asymmetry across the five providers). No major AI chatbot provider currently publishes aggregate statistics on the number of conversations referred to law enforcement for threats of violence, terrorism, or serious harm to others.⁶⁰ The Tumbler Ridge, FSU, and Strasbourg cases are known only because they became public through tragedy, litigation, or press reporting. It is reasonable to assume that other referrals have

⁵⁷ *United States v. Dost*, 636 F. Supp. 828, 832 (S.D. Cal. 1986), *aff'd sub nom. United States v. Wiegand*, 812 F.2d 1239 (9th Cir. 1987).

⁵⁸ On the rise of AI-generated and synthetic CSAM and the corresponding stress on hash-matching infrastructures, see among many other sources [NCMEC's 2024 CyberTipline Report](#), which records a sharp year-on-year increase in reports involving generative-AI-produced material.

⁵⁹ The chapter's particular focus on provider-initiated referrals based on violent-threat detection is driven by the public episodes examined in Section 1.2 (Tumbler Ridge, FSU, Strasbourg), which have surfaced this category in a uniquely concrete way. Other categories of provider-initiated referrals (financial crime, fraud, terrorism unrelated to imminent violence, suicide and self-harm escalations involving identifiable third parties) raise comparable transparency questions; the analytical points made in this Chapter in relation to violent-threat referrals apply, with appropriate adjustments, to those neighbouring categories as well.

⁶⁰ Some non-AI platforms publish related figures (Twitch's transparency reports, for instance, include data on emergency-disclosure responses to law enforcement), and the absence of equivalent reporting in the AI chatbot setting is therefore a gap relative both to existing CSAM reporting practice and to comparable practice in adjacent online services.

occurred without public knowledge, but it is impossible to assess their frequency, the thresholds applied, the jurisdictions involved, or the outcomes. A governance mechanism that can lead to armed police intervention should, at minimum, be accompanied by periodic aggregate reporting: the number of referrals made, the general categories of threat involved, and the jurisdictions to which disclosures were directed. Because the proposal is for aggregate figures rather than case-level disclosure, it presses for visibility without exposing the individual referral decisions that, as Section 1.4 notes, a litigant could mine as evidence of knowledge; it therefore does not sharpen the documentation paradox that might otherwise discourage referral. CSAM reporting itself does not provide every element of this proposed standard (it typically includes the number of reports and the aggregate volume of items, without the categorical breakdowns and jurisdictional information suggested above), but it establishes a baseline of public quantitative reporting that the violent-threat category does not currently match. There is no principled reason why violent-threat and other emergency referrals should not at least reach that baseline, and the case set out above for adding the categorical and jurisdictional dimensions would strengthen accountability further across both domains.

The case for such reporting rests on three substantive grounds:

First, it provides public visibility into the operational application of thresholds that, as Section 1.1 sets out, currently lie almost entirely with private providers; without aggregate data, the actual operating thresholds are inferable only from the small number of referrals that reach the public record through tragedy, litigation, or press attention.

Second, it generates comparative data across providers, which is the precondition for any informed policy debate about whether thresholds are well-calibrated and whether they are converging or diverging across the industry.

Third, it provides statistical context for individual cases, which in its absence must shoulder the entire empirical weight of the public discussion (and which, as the discussion of Tumbler Ridge in Section 1.2 illustrates, can therefore be made to bear more analytical weight than three or four cases can really support).

1.6. The tension with cryptographic hardening and Sealed Mode

The analysis of provider-initiated disclosure creates a direct and, in important respects, unavoidable tension with the Sealed Mode framework proposed in Part 1 of this study.

Sealed Mode's sixth component, cryptographic hardening, envisions, for some sensitive conversations, strengthened technical protections that move the confidentiality boundary from one enforced by policy commitments to one enforced, to the extent feasible, by architecture. But several of the episodes analysed in this chapter illustrate a feature of provider-initiated disclosure that bears directly on cryptographic hardening: in many cases (and certainly in the Tumbler Ridge and Strasbourg episodes), the underlying detection relied on the provider's ability to access conversation content in plaintext.

Not all provider-initiated disclosures operate this way (a meaningful share of emergency disclosures across the technology sector arise from external tips, reports from other users, information from law-enforcement counterparts, or signals derived from non-content logs and metadata rather than from the provider's own content monitoring), and post-incident cooperation similarly draws on a mix of content records, non-content logs, account metadata, and information obtained through other channels.

The point is therefore narrower than a categorical claim that proactive disclosure requires content access: it is that the specific mechanism of content-based detection and referral, which several of the cases in Section 1.2 illustrate, depends on plaintext access in a way that complete cryptographic hardening would foreclose. None of the three cases analysed above concerned health-related or other sensitive conversations of the type intended to be covered by Sealed Mode; the Garcia litigation discussed in Section 1.4, however, illustrates that mental-health conversations can themselves become life-threatening, which means the tension between cryptographic protection and safety monitoring could in principle run through Sealed Mode's core use case as well, unless providers introduce adequate

mitigations (such as refusal to engage with certain categories of life-threatening content) within the conversational interface itself.⁶¹

This tension is not unique to chatbots. It is a specific manifestation of the broader debate between end-to-end encryption and lawful access that has shaped communications policy for two decades, from the original Crypto Wars of the 1990s to the post-9/11 surveillance debate, the 2015-2016 FBI–Apple confrontation over the San Bernardino iPhone, the European debate around proposed CSAM-scanning regulation, and the periodic resurgence in the United States of legislative proposals that would condition intermediary liability protections on content-scanning practices.⁶²

The chatbot context, however, sharpens the tension in a particular way. The sensitivity that the providers themselves attribute to these conversations, set out in the Introduction, is what makes the case for strong confidentiality and the case for content-based safety monitoring compelling at once: the same material users most want protected is also the material whose monitoring can avert harm. The user who types “I am terrified it might be something serious” deserves privacy protections commensurate with the trust they place in the system; the user who describes an operational plan to kill someone may generate a conversation that, if not flagged, leads to catastrophe.

Whether this tension can be partially resolved through design or whether it represents a genuine binary choice is itself debated. Peter Swire has argued, in the context of end-to-end encryption, that one must sometimes choose between two outcomes: weak security where safety scanning operates, or strong security where it does not.⁶³ If that analysis is correct, then the tension described here is not a design problem to be solved but a policy choice to be made. The choice would be between:

- (a) A mode in which the provider genuinely cannot access conversation content, meaning that content-based safety monitoring is blind and content-based proactive disclosure is foreclosed, while other disclosure pathways (those based on non-content signals, external reports, or post-incident reconstruction from metadata and logs) continue to operate; and
- (b) A mode in which content-level access is preserved, enabling content-based safety monitoring but also exposing conversations to the full range of access risks documented in this study and in Part 1.

This chapter does not resolve that choice. It identifies it as one of the most important questions that providers, regulators, and the technical community will need to confront. The choice may not, however, be entirely binary in practice. Some technical architectures, such as differential-privacy techniques applied to aggregated safety signals, could in principle preserve some safety-monitoring capacity while substantially constraining the provider’s access to plaintext. Whether such intermediate architectures are technically feasible, sufficiently reliable, and compatible with existing legal obligations is an open question that deserves sustained research attention.

There is also a contextual dimension that has received insufficient attention. The appropriate balance between cryptographic hardening and content-based safety monitoring may differ significantly depending on the environment in which the chatbot operates. For a general-purpose consumer chatbot such as ChatGPT or Claude, accessible to hundreds of millions of users through a dedicated interface, the cases analysed in this chapter (Tumbler Ridge, FSU, Strasbourg...) illustrate that content-based safety monitoring has, in concrete instances, produced referrals that led to armed intervention, criminal investigation, and (on some readings of the same evidence) the possibility of preventing further harm. Removing that monitoring capacity entirely through complete cryptographic hardening would be a decision that legislators, regulators, and the public would need to debate openly.

The calculus shifts when chatbots are integrated into messaging services that are already built on end-to-end encryption. Meta AI is already embedded in WhatsApp, and if other AI assistants follow into platforms such as Signal, they enter environments whose security architecture was deliberately designed to prevent the provider from accessing message content. The technical interaction warrants being made

⁶¹ On the Garcia litigation and the broader chatbot-related civil-litigation wave, see Section 1.4 above.

⁶² For the broader debate, see Berkman Klein Center for Internet and Society at Harvard University, “[Don’t Panic: Making Progress on the Going Dark Debate](#)”, 1 February 2016. The literature is now extensive; on more recent developments in the EU context, see also the ongoing debate around the proposed Regulation laying down rules to prevent and combat child sexual abuse (the “CSAM Regulation”), and on the US side the periodic resurgence of the EARN IT Act and related legislative proposals.

⁶³ Peter Swire’s intervention in “[Whispers to Machines: When Chats Become Evidence](#)”, Privacy Symposium Panel, 21 April 2026, Venice.

explicit. WhatsApp messages between users are end-to-end encrypted, so that neither WhatsApp nor Meta has access to their content in transit or at rest on Meta's servers. When a user addresses a query to Meta AI within WhatsApp, however, that query leaves the end-to-end encrypted pathway and reaches Meta's AI servers in a form readable to Meta's infrastructure, because the AI provider's processing requires it; the response is generated and returned to the user on the same basis. An AI assistant embedded in an end-to-end encrypted messaging app can therefore, depending on how it has been engineered, create a non-encrypted pathway through an otherwise encrypted system.

As this study went to press, this concern has been materially addressed for the leading example. On 13 May 2026, Meta announced *Incognito Chat with Meta AI*, extending its existing Private Processing architecture from peripheral AI features to the chat conversation itself between a user and Meta AI inside WhatsApp. The architectural anomaly described in the preceding sentence is precisely what the announcement is intended to close: the chat is now routed through a Trusted Execution Environment in which the provider, on the public description of the architecture, **cannot read its content**. A detailed analysis of the announcement and its implications for the framework set out in this chapter is offered in a dedicated paper with Peter Swire;⁶⁴ the conclusions to this study return to its broader significance.

The substantive concern this raises is not best framed in terms of user expectations. User expectations are themselves shaped by how products are designed: a user familiar with the way Meta AI operates within WhatsApp would, over time, develop a different set of expectations from a user accustomed to a hypothetical AI assistant engineered for end-to-end behaviour from the outset, and the appeal to expectations therefore cannot do the normative work here without circularity. The substantive concern is that the policy and architectural commitments instantiated by an end-to-end encrypted messaging platform have weight in their own right, independently of how any particular user's expectations were formed. An end-to-end encrypted platform is a deliberately instantiated choice to prioritise confidentiality over content-based monitoring; that choice has been accepted by users, by regulators, and by platform operators over many years, and it is presupposed by the surrounding legal and policy environment in which the platform sits (including its treatment under data-protection law, its relationship to law-enforcement access regimes, and the public commitments it has made to its user base). Embedding an AI assistant into such a platform in a way that bypasses that prior choice (by routing queries in plaintext through provider servers, by retaining records on the provider side, or by enabling content-based safety monitoring of conversations that would otherwise be inaccessible to the provider) materially weakens the platform's overall security posture without necessarily being visible to its users. That is a substantive design and policy concern, distinct from any claim about what users do or do not subjectively expect.

The point is best put in design terms rather than expectational ones. The level of cryptographic protection appropriate for chatbot conversations depends on the architectural environment in which the chatbot is deployed, because that environment carries prior commitments that the chatbot inherits. The governance framework for cryptographic hardening should therefore be contextual rather than uniform, distinguishing at minimum between standalone consumer chatbots (where some form of safety monitoring may be justified subject to the accountability questions examined in Section 1.7 below), chatbots embedded in end-to-end encrypted environments (where coherence with the surrounding security architecture is the natural default, and any departure from it warrants explicit justification and user disclosure), and enterprise or health-specific chatbots operating under Sealed Mode (where the appropriate level of protection will depend on the specific regulatory and clinical context).

The tensions extend beyond cryptographic hardening to a second Sealed Mode component: strict retention limits. Part 1 documented that reviewed conversations can be retained for up to three years after deletion and that safety-classification scores may persist for up to seven years. The retention question interacts with both proactive and post-incident disclosure, in ways the chapter's earlier analysis suggested might be more separable than they actually are. Proactive disclosure based on real-time detection (the Strasbourg flagging is one example) operates on current conversational material, but a provider that detects a concerning conversation in real time may also look back at the user's historical

⁶⁴ See Theodore Christakis and Peter Swire, "[The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI](#)", AI-Regulation Papers 2026-05-4, May 2026.

conversations, where they have been retained, to inform the referral decision. Post-incident cooperation is more clearly oriented to retained material.⁶⁵ The retention question is therefore relevant for both phases of provider-initiated disclosure.

The Sealed Mode framework must nevertheless confront an uncomfortable internal tension: the same retention practices that Part 1 identified as privacy risks are, in the proactive-disclosure context, evidentiary and investigative resources. The resolution cannot be to abandon retention limits altogether, which would leave the privacy concerns documented in Part 1 unaddressed. Nor can it be to enforce aggressive deletion schedules without acknowledging that doing so may reduce the evidentiary base on which post-incident cooperation, and the look-back component of real-time proactive disclosure, currently draw. The most promising direction is a tiered approach: strict retention limits as the default for ordinary conversations, with a narrow, clearly defined, and auditable exception for conversations that have been flagged by safety systems, where retention serves a documented law-enforcement or child-safety purpose. That exception must itself be subject to time limits, oversight, and transparency reporting of the kind argued for in Section 1.5 above.

As Peter Swire has observed, there is no legal duty of comprehensive surveillance in the United States or the European Union; under EU law, data minimisation and data-protection-by-design principles strongly push back against any such obligation.⁶⁶ The issue is therefore not whether providers must monitor everything, but what happens when they *do* monitor (as all the providers examined in this study currently do for safety purposes) and what legal and institutional framework governs the resulting referral decisions and the retention practices that support them.

1.7. Lessons from social media moderation, Google search, and human confidants: why chatbots raise distinctive accountability questions

The fundamental question raised by provider-initiated referrals (decisions made by internal corporate teams without external authorisation or independent review) is not novel in form. Private companies have made decisions of this kind, across many adjacent settings, for decades, and the institutional debate around such decisions has its own substantial literature.

What the chatbot context adds is not the bare fact of private decision-making but a specific combination of features that the three comparisons below are designed to bring into focus: with social media moderation, which is the closest institutional analogue and the most extensively analysed; with general-purpose search, which illustrates how a comparably large-scale information service has publicly framed its law-enforcement posture; and with the human confidants in whose register chatbot users actually speak, which is rarely treated in this literature but which most directly illuminates what is genuinely distinctive about the chatbot setting.

Social media moderation

Social media platforms have faced the fundamental question of private moderation and external referral for over a decade. When a platform decides to remove content, suspend an account, or refer material to law enforcement, that judgment is made by private employees applying private policies, without judicial authorisation or independent review. The resulting debate has generated a substantial body of scholarship, litigation, and institutional experimentation, including the Meta Oversight Board's attempt to introduce quasi-independent appellate review of content-moderation decisions.⁶⁷ That institutional

⁶⁵ But here too the picture is more mixed than it first appears: while stored content is often important, non-content logs, account metadata, and information obtained through other channels (including external reports and inter-provider cooperation) are frequently relevant alongside, or in place of, stored content.

⁶⁶ Peter Swire's intervention in "[Whispers to Machines: When Chats Become Evidence](#)", Privacy Symposium Panel, 21 April 2026, Venice.

⁶⁷ The Meta Oversight Board, established in 2020, provides independent appellate review of selected Meta content-moderation decisions. The literature on its institutional design and effectiveness is now substantial; see, for an overview, the [Oversight Board's annual reports](#). The Board is not a model that translates directly to provider-initiated law-enforcement referrals (its remit is content moderation, not external disclosure), but its existence is the most prominent institutional experiment in the social-media accountability space and is therefore the natural reference point for the comparative analysis that follows.

history is the natural reference point for any analysis of accountability in the chatbot context, and several of its lessons (the value of aggregate transparency reporting, the difficulty of designing review mechanisms that operate at scale without becoming bottlenecks, the limits of self-regulation in the absence of external pressure...) carry over straightforwardly.

The chatbot context nevertheless **differs** from traditional social-media moderation in at least two respects that bear on the accountability analysis.

First, the nature of the underlying material differs, and the difference is fundamental. The relevant comparison is with public-facing social-media posting, the characteristic and, on platforms such as X and LinkedIn, overwhelmingly dominant way these services are used; it is not with the private-messaging features that some platforms also offer in parallel, which carry the ordinary confidentiality of any private communication and lie outside this comparison.

A post, in the sense at issue, is content the user has chosen to address to an audience, whether the general public or a defined circle of contacts ("friends"). A chatbot conversation is addressed to no audience at all; it is a private exchange conducted, as the Introduction sets out, in a register closer to the conversations one sometimes has with a therapist, a spouse or a close friend. The moderation and referral apparatus of social-media platforms, and much of the law around it (defamation, hate-speech regulation, intermediary liability under the DSA and equivalent regimes), grew up precisely because the content was externalised to others.

Content addressed to an audience therefore carries a different, and generally lower, expectation of confidentiality than content addressed to no one but the system itself, and for a wholly public social media post that anyone in the world can read that expectation is effectively non-existent. The distinction is one of kind, not of degree. A social-media post is, by definition, addressed to others: to the public at large, in the open posting that dominates platforms such as X and LinkedIn, or to a defined circle, in the case of a friends-limited post. A chatbot conversation is addressed to no audience at all. The user who writes to a chatbot is not publishing to anyone; in the user's perception, and in the way the interface invites the service to be used, the exchange is a private one. A friends-limited post sits between the two, retaining some expectation of privacy, greater than a wholly public post and less than a private exchange, but it remains addressed to an audience, the very feature a chatbot conversation lacks. Whether the user's expectation of confidentiality is in fact matched by the legal and architectural reality is the question this study exists to examine; the point here is only that the expectation is of a different order from the one a social-media post can support, and that a referral of such a conversation to law enforcement is an intrusion of a kind the social-media moderation apparatus was not built to handle.⁶⁸

Second, the interpretive challenge differs. Social-media moderation typically evaluates discrete, self-contained statements: a single post, image, or video. Chatbot conversations are often extended, contextual, and ambiguous in a way that single posts rarely are. A user who types "I want to kill my boss" may be venting, role-playing, expressing frustration as one might to a spouse or a colleague at the end of a hard day, testing the system, drafting fiction, or expressing genuine intent. The same words in a social-media post carry different contextual signals from those carried by the fifteenth message of a late-night conversation with a system that has been designed to respond with empathy and engagement. The interpretive challenge is correspondingly greater, and the consequences of a misjudgement, whether the failure lies upstream in the misinterpretation of the conversational material or downstream in inadequate verification by the receiving authority once an alert has been issued (see Section 1.2 above), can be severe.

Google search

⁶⁸ There is a further distinction worth noting. Proactive referral of social-media content to law enforcement engages free-speech concerns directly: the content referred is communication addressed to an audience, and the First Amendment in the United States and Article 11 of the EU Charter protect the right to express views publicly, even controversial ones, against government sanction. Chatbot conversations do not raise the same free-speech concern in the traditional sense, because they are not addressed to a public audience. But the chilling effect may be no less serious. If users come to understand that their private conversations with an AI system are monitored and may be referred to law enforcement, the inhibition extends not to public expression but to inner thought: the exploratory reasoning, emotional processing, and speculative questioning that people engage in precisely because they believe they are in a private space. The social cost of chilling public speech is well documented and constitutionally recognised. The social cost of chilling private thought is less studied but may ultimately be more damaging to the distinctive value that conversational AI provides.

A second comparison, with general-purpose search engines, is of a different kind from the social-media one. Where that comparison turned on the nature of the underlying material, this one turns on public posture: how a comparably large-scale information service has publicly framed its cooperation with law enforcement, and how the chatbot industry's more proactive stance compares.

Google Search processes very large volumes of queries daily, and providers in this space have therefore long had to think about what to do with queries whose content would attract attention in other settings. Google has published policy positions on these questions. On the basis of Google's transparency materials, Google reports apparent child sexual abuse material to NCMEC when it becomes aware of such material;⁶⁹ for other categories of content, Google's public law-enforcement posture is framed primarily around responding to government requests through established legal process (subpoenas, court orders, warrants),⁷⁰ and Google states that, in emergencies in which it "reasonably believe[s] that [it] can prevent someone from dying or from suffering serious physical harm", it "may provide information to a government agency, for example, in the case of bomb threats, school shootings, kidnappings, suicide prevention, and missing persons cases".⁷¹

What is observable from the public record is more limited than a confident claim about industry practice would require, and the point should be made carefully. Google does not publicly describe operating a general system for proactively scanning alarming search queries and referring users to law enforcement on that basis outside the recognised CSAM and emergency frameworks. Whether such a system operates internally is not knowable from the public record, and the absence of public reporting cannot be taken to mean that no proactive referrals based on search-query content occur. What can be observed is that Google's publicly described posture frames its emergency-disclosure cooperation in a particular way: as a response to government requests (including emergency requests), rather than as the output of a Google-initiated screening pipeline. The publicly described posture of major consumer chatbot providers is different in emphasis: OpenAI, in particular, has described a pipeline in which automated systems identify potentially concerning activity at scale, trained personnel review flagged accounts or conversations in context, higher-risk cases are escalated for more detailed investigation, and conversations indicating an imminent and credible risk of harm to others may be notified to law enforcement on the provider's initiative (see Section 1.3 above).

The conclusion to draw from this descriptive contrast is more limited than a strong claim about industry practice would warrant. It is not that one industry refers more often than the other, which is not knowable from public information. It is that the chatbot industry has publicly committed to a more visible proactive role in detecting and referring concerning content than the search industry has done, and the policy and accountability questions about that role are correspondingly more salient in the chatbot setting. The point is genuinely descriptive, not normative: the comparison illuminates what a major large-scale information service has publicly chosen NOT to position itself as doing, against which the chatbot industry's public posture can be analytically situated.

Human confidants

A third comparison, less often made in this literature but particularly illuminating, is with the human confidants in whose register chatbot users actually speak. The case for this comparison is empirical rather than analogical. As Section 1.1 sets out, the buddy framing of consumer chatbot use is not a marketing slogan but the mode in which the use is conducted at scale: users confide in chatbots as they would in a spouse, a close friend, a parent, or a therapist, in a register that they would not generally

⁶⁹ Google, "[Google's Efforts to Combat Online Child Sexual Abuse Material FAQs](#)", Transparency Report Help Center ("Once Google becomes aware of apparent CSAM, we make a report to NCMEC").

⁷⁰ Google, "[How Google handles government requests for user information](#)", Privacy & Terms (distinguishing between subpoenas, court orders, and search warrants); see also Google Transparency Report, "[Requests for User Information](#)".

⁷¹ Google, "How Google handles government requests for user information", note above. On the basis of the public record, Google's emergency-disclosure framework operates primarily in response to government emergency requests rather than as a Google-initiated screening pipeline. A law-enforcement agency that has become aware of an emergency through its own investigative channels (a witness report, a missing-person investigation, an intercepted threat) asks Google for information held in a specific account or associated with a specific search query, and Google evaluates the request. See Google, "[Requests for User Information FAQs](#)", Transparency Report Help Center ("Where a person may be in serious physical danger, a government agency may ask Google to voluntarily disclose information needed to prevent the emergency").

adopt in writing to a human counterpart. The comparison is therefore not with a hypothetical analogue but with the actual relationships whose register chatbot conversations most resemble.

Consider the same utterance across these settings. When someone says to a spouse, a close friend, or a parent at the end of a hard day that they “want to kill the boss”, this is not a referral occasion. The listener has continuing context that allows them to read what the utterance is doing (it is venting, processing, or seeking validation, almost never planning); they have no legal duty to refer it to the authorities; and the system in which the speech takes place is protected by spousal privilege (for spouses) or by social and cultural norms of confidentiality (for other intimate relationships) that limit the compelled disclosure of such communications. When the same utterance is directed to a therapist, an even more structured set of protections applies: psychotherapist-patient privilege limits the compelled disclosure of the communication; the therapist has clinical training to assess threats and a continuing professional relationship through which to read what a particular utterance is doing; and the duty to warn that *Tarasoff* established (see Section 1.1 above) applies only in the narrow circumstances in which an identifiable third party is at credible and imminent risk.⁷² In each case, the legal-institutional architecture surrounding the speech is calibrated to the relationship in which it takes place: confided speech receives confided treatment, with narrowly defined exceptions for grave and identifiable risks.

The chatbot setting receives speech in the same register but supplies almost none of the legal-institutional architecture that has historically attended such speech. A user who confides in a chatbot enjoys none of the evidentiary privileges that attach to communications with a spouse or a therapist, and none of the social and professional confidentiality norms that attend intimate human relationships. This is not to say the communication is legally unprotected: it carries, in the USA, statutory protection under the Stored Communications Act and its Fourth Amendment status against compelled government access is a live and contested question, as Chapter 2 discusses. What it lacks is the privilege and the relational confidentiality that have historically attended speech in this register. The provider has no professional duty of care of the kind that constrains therapists; there is no continuing relationship of the kind that allows a human listener to read what a particular utterance is doing (memory features in some chatbots provide conversational history, but not the relational context the listener-as-confidant relationship supplies); and the provider has the operational capability, and (under § 2702) the permission, to refer the communication to law enforcement based on a good-faith assessment that an emergency exists. The issue is not that chatbots create speech that did not exist before; people have always confided in human confidants, including in alarming registers. The problem is that the speech that has historically been protected by privilege and contextual interpretation is now being received by a system that supplies neither, and that is also institutionally configured to refer such speech upward through an internal pipeline whose criteria are not publicly visible.

This observation has a normative implication that the chapter should be explicit about. Two forces pull in the same direction here. People disclose intimate things to those whose presence invites confidence, close relatives, friends and clinical professionals, often things they would not put in writing to a wider audience. But people also disclose intimate things precisely because the listener is not someone who can judge them: a behavioural literature on self-disclosure, discussed in the Introduction, finds that the perceived anonymity and reduced fear of judgement of a non-human interlocutor can prompt disclosures a person might withhold from a friend or a spouse. Chatbot interfaces are designed in ways that elicit precisely this inclination, and along both routes: responsiveness, an empathetic register and continuous availability close the social distance between speaker and listener, while the felt absence of a human who might judge lowers the threshold of what a user will say.⁷³ The “buddy” framing that arises is therefore not a mere accidental misreading of the technology by naïve users.

The provider that designs such an interface bears a corresponding responsibility, distinct from the referral-threshold question developed earlier in this section, to be clear with users about the conditions under which they are speaking: that conversations are recorded, are subject to automated review, may

⁷² On the *Tarasoff* doctrine and its structural distance from the chatbot context, see Section 1.1 above. On spousal privilege, the federal rule in the United States is set out in *Trammel v. United States*, 445 U.S. 40 (1980) (recognising both the testimonial privilege and the confidential-communications privilege); state-law formulations vary. On psychotherapist-patient privilege under federal law, see *Jaffee v. Redmond*, 518 U.S. 1 (1996).

⁷³ More restrained, businesslike configurations are possible and are used by some, particularly expert users, but they are not the default mode in which consumer chatbots are encountered at scale, which is the relevant mode for a study of consumer use.

in defined circumstances be referred to law enforcement, and do not come with the privileges, professional duties of care or confidentiality norms that surround the human confidant relationships in whose register users have come to speak. The point is not that chatbots should not be used in this way, since people will and there is value in it; it is that users should be able to understand the terms of an exchange that the design itself tends to obscure. Nor is this the generic proposition that companies should disclose their data practices, true though that is. It is the narrower duty to be honest about the one thing the confiding register invites a user to assume and that is not in fact present: the protective architecture of a genuine confidant relationship. There is also a more demanding version of the same point that some commentators are beginning to make: that interface design choices which actively encourage the “buddy” framing (warmth cues, anthropomorphic identity, continuity of conversational persona) may themselves require justification when the legal-institutional architecture surrounding the speech they elicit is so far from what the framing suggests.

What is and is not new

None of the individual components is new, and it is worth saying so plainly. Provider emergency disclosure under § 2702 has operated for decades across email, messaging and cloud storage; private moderation decisions affecting users have been studied for fifteen years; cross-border information-sharing chains between providers and authorities are familiar from financial-intelligence and counter-terrorism work. A practitioner from any of those fields will recognise the institutional questions raised here in their general form, and will be right to.

What is new is the conjunction, and it is not a subtle one. A large and growing share of these conversations, hundreds of millions of them, is conducted in the register people often reserve for a therapist, a doctor, a spouse or their closest friend; and that speech now enters systems that record it, subject it to automated review, and can refer it to the police on the provider’s own initiative, while carrying none of the privilege, professional duty or confidentiality norm that has always surrounded speech in that register.

What distinguishes the chatbot, taken one comparison at a time, is this. Against social media, the speech is confided rather than addressed to an audience. Against general-purpose search, it is sustained and contextual rather than a one-off query, an utterance embedded in a conversation long enough to supply interpretive material that a single search does not. And against the human confidant, whose intimate register these conversations borrow, it supplies none of the privilege, professional duties of care, confidentiality norms or continuing relationship that the human relationship carries. No one of these features is new on its own; what is without precedent is their convergence on a single, vast category of intimate communication, and that is what warrants the analytical attention this chapter has given it

The accountability question, properly framed

This analysis bears directly on how the institutional question should be framed, and the framing should be made explicit. The chapter has not been arguing, and should not be read as arguing, that an institutional gap exists in the sense that the police lack accountability for what they do with provider referrals. Police accountability operates through familiar mechanisms: warrants required for searches and seizures, civilian oversight bodies, internal-affairs investigations, civil-rights litigation, prosecutorial review of charging decisions, and the political accountability of elected officials. When a provider refers a conversation to law enforcement, the receiving authority makes its own assessment of the credibility and severity of the threat, gathers its own evidence through its own investigative tools, and makes its own decisions about how to respond, subject to the accountability mechanisms that govern policing generally. The discussion in Section 1.2 of the receiving authority’s independent role makes the same point in a different register.

What the chapter does identify is a different and narrower question. The criteria by which providers determine which conversations to refer in the first place are not publicly visible, are not subject to any structured external review, and are not currently the subject of systematic transparency reporting beyond the CSAM domain. This is a question about the inputs to the police-accountability system, not about the police-accountability system itself. The analogous question in the social-media context (the

criteria by which platforms decide which content to remove or to refer) has been the subject of more than a decade of sustained policy debate, has produced institutional experiments such as the Oversight Board, and remains contested; the analogous question in the chatbot context has only recently begun to receive equivalent attention. The chapter's analytical claim is best understood not as a gap-in-accountability claim but as **a transparency-of-threshold-setting claim**: the absence of public information about how providers set referral thresholds, against the institutional backdrop set out in the three comparisons above, is the specific feature of the current landscape that warrants attention.

What is, and is not, being proposed

Three things follow from this framing that are worth being explicit about. The chapter is not proposing in any way that mandatory reporting laws be imposed on chatbot providers in the violent-threat or equivalent-emergency categories. The chapter is not proposing either that referral decisions be subject to pre-authorisation by judicial or quasi-judicial bodies before the fact: such authorisation would be impractical in genuine emergencies and would create operational pathologies that the cases in Section 1.2 have already begun to illustrate. The chapter is not proposing a single institutional design.

What the chapter is proposing is more modest, and is set out in the recommendations that follow in Section 1.8. In summary form: that major chatbot providers publish aggregate statistics on provider-initiated referrals in violent-threat and equivalent emergency categories, on the model of (and at minimum at the baseline established by) existing CSAM reporting practice; that the internal criteria applied to such referral decisions, while not necessarily themselves publishable in full, be sufficiently documented internally; that periodic independent audit of their application becomes possible (the conditions under which such audit would be conducted, and by whom, being themselves questions that warrant further work). The chapter has also argued, in Section 1.2 above, that an accountability framework focused only on providers misses the receiving authority's investigative role, and any institutional design that follows from the recommendations above should be calibrated with both stages in mind. Beyond these baseline measures, the chapter calls for the sustained attention of legislators, regulators, experienced litigators, and institutions such as the Sedona Conference that have expertise in developing practical frameworks for complex disclosure problems.

Litigation as a parallel transparency pathway

The civil and criminal litigation now accumulating against major chatbot providers, documented in Section 1.4 above, may paradoxically become an important near-term source of public information on how chatbot providers actually make provider-initiated disclosure decisions. Voluntary corporate commitments have so far disclosed almost none of the operative details: the criteria applied, the escalation paths followed, the role of human reviewers, the conversations referred and the conversations not referred. Civil discovery, criminal subpoenas, and adversarial pleadings have begun to extract those details against the providers' will.

If these cases proceed to discovery, as several appear likely to do, the resulting process could produce significant new information, likely under protective orders given the confidential trade-secret dimensions involved. If any of them proceeds to trial, which remains comparatively rare in this kind of litigation, the evidence would likely be presented in open court. Either way, the litigation pathway could yield disclosure of the internal criteria that major providers apply when deciding whether to refer a flagged conversation, the error rates of their automated classifiers, the role of human reviewers in the escalation chain, the specific conversations that led to or did not lead to referrals, and the corporate deliberations that shaped the referral threshold before and after specific incidents. The transparency that voluntary reporting has not yet produced may thus arrive through adversarial litigation, a path that is neither efficient nor desirable, but that may prove to be an important near-term source of public information about how provider-initiated disclosure decisions are actually made.

Table 1. Comparative overview of proactive corporate disclosure

Note: Sections 1.1 through 1.7 make it possible to set out, in comparative form, what is currently publicly known about how each of the five major consumer chatbot providers examined in this chapter approaches proactive disclosure. Table 1 below summarises that comparison across the dimensions that the preceding analysis has identified: the public basis for proactive disclosure, the publicly documented detection mechanisms, the strongest documented reporting category, the main transparency gap, and the publicly stated approach to user notification after a proactive disclosure has been made. The table is descriptive: it captures what the public record shows, not what may be operating internally that the public record does not reflect. Assessments based on publicly available materials as of 8 April 2026.

Dimension	ChatGPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	Grok (xAI)	DeepSeek
Public basis for proactive disclosure	Emergency disclosure for danger of death or serious physical injury; broadened referral criteria confirmed in February 2026 Canadian letter. Retroactive cooperation confirmed in FSU case (April 2025).	Emergency-disclosure process and child-safety reporting are public; routine violent-threat practice remains less visible.	Google-wide emergency-disclosure and child-safety policies apply; no Gemini-specific referral report located.	Public CSAM reporting commitment and general legal-safety disclosure language; no dedicated violent-threat framework located.	Privacy policy allows good-faith emergency and legal-compliance sharing; no dedicated proactive-referral framework located.
Publicly documented detection mechanisms	Usage monitoring, auxiliary models, blocking, human review, enforcement, and law-enforcement escalation channels (Preparedness Framework).	Hash matching, detection classifiers, and trust-and-safety review.	Automated child-safety tooling, classifiers, and human review across Google services.	Public NCMEC reporting commitment; little public detail on detection architecture.	Safety monitoring mentioned in privacy policy; technical detection architecture not publicly detailed.
Strongest documented reporting category	CSAM (107,817 CyberTipline reports, H2 2025); emergency serious-harm referral architecture publicly acknowledged but not statistically reported.	CSAM (5,005 pieces reported to NCMEC, H2 2025); emergency disclosure policy publicly stated but violent-threat metrics not reported.	CSAM/child sexual exploitation reporting across Google services; no Gemini-specific proactive-referral metrics.	Suspected CSAM.	Emergency protection of life or health and legal-compliance sharing.
Main transparency gap	No regular public statistics for violent-threat referrals; no public account of referral decision criteria beyond “credible and imminent” (now broadened).	No provider-specific metrics for violent-threat referrals.	No Gemini-specific proactive-referral metrics.	No Grok-specific transparency report on proactive referrals.	No dedicated public transparency report located.
User notification after proactive disclosure	No provider has published a specific user-notification policy for proactive voluntary referrals as distinct from compelled disclosure	No provider has published a specific user-notification policy for proactive voluntary referrals as distinct from compelled disclosure	No provider has published a specific user-notification policy for proactive voluntary referrals as distinct from compelled disclosure	No provider has published a specific user-notification policy for proactive voluntary referrals as distinct from compelled disclosure	No specific user notification policy located.

1.8. Conclusions and tentative recommendations

The governance question this chapter has examined is genuinely difficult. It involves competing legitimate interests (preventing imminent violence or self-harm; preserving a space for candid, exploratory, and at times difficult conversation), competing institutional realities (the operational demands of moderation at scale; the procedural safeguards -warrants, judicial oversight, due process...- that ordinarily protect citizens in their interactions with the state), and competing temporal pressures (the speed required to act on credible threats; the deliberation required to avoid disproportionate intrusions). The chapter does not propose a general resolution. What it offers is an empirical mapping of how the practice now operates, an analytical account of what makes the chatbot case distinctive, and a set of tentative recommendations that build on existing institutional infrastructure rather than starting from scratch.

The cases of Tumbler Ridge and FSU, examined in Sections 1.2 and 1.4, may already represent for the chatbot industry the beginnings of what the Christchurch attack of March 2019 represented for social-media moderation: a category of incident that catalyses regulatory and reputational change. The comparison has its limits. The Christchurch attack killed 51 people in a single live-streamed event whose impact on platform regulation was bound up with its scale and its visual specificity, and neither of those features translates straightforwardly to the chatbot setting. What the comparison nevertheless brings into focus is the following pattern: regulatory and reputational architecture in a digital sector tends to emerge in the wake of specific catalysing events. The cases in the chatbot setting are accumulating; the public-record evidence of provider-initiated emergency referrals is rapidly increasing; and the chatbot industry, the regulators that supervise it, and the academic and practitioner communities that study it have an opportunity to do the analytical and institutional work in tandem with the events that are unfolding. The recommendations that follow are offered in that spirit.

Much of the framework this chapter analyses is not new in its general form. Emergency disclosure by electronic communication service providers under good-faith standards has operated for decades; trust-and-safety functions at major providers routinely consult academic experts, clinical specialists, and civil-society actors; self-harm escalation predates the current chatbot generation. What is distinctive about the chatbot setting is not the bare fact of provider-initiated disclosure but a specific combination of features that does not appear in any of the adjacent settings taken individually. Against the shared backdrop of automated processing at scale, which the chatbot shares with search and social media alike, three features set it apart: the confided register in which a large and growing share of consumer chatbot use is conducted (users often speak to chatbots in ways they would not adopt in writing to a human counterpart, in registers more reminiscent of conversations with a therapist, a doctor, a spouse or a close friend than of public posts or search queries); the conversational continuity and interpretive ambiguity that surround individual utterances; and the absence of the confidentiality norms, professional duties of care and continuing relational context that have historically surrounded such speech in adjacent human relationships. The novelty is in the conjunction, not in any element on its own.

Three tensions frame the governance agenda this chapter identifies. They reflect competing considerations that must be balanced, and each is best illustrated by a concrete example.

The first tension is between user safety and conversational confidentiality. Consider a user who confides in a chatbot, late at night, that they have been thinking about ending their life. The safety interest is in routing the user, quickly, to appropriate support, and (if the risk appears imminent) to emergency services. The confidentiality interest is in not building a system in which a candid disclosure of distress triggers contact with the state. These interests can coincide: a privacy-respecting system may better serve safety by preserving users' willingness to seek help through legitimate channels rather than retreating from systems they no longer trust. They can also diverge: an automated classifier may flag the same conversation that a clinical reader would treat as a non-emergency expression of distress, with the result that emergency services are contacted (a police welfare check, or a comparable intervention) in a case where such contact would not have been judged necessary by a human reviewer with the conversational context in hand. The legal frameworks for emergency disclosure⁷⁴ attempt a calibrated answer to this

⁷⁴ 18 U.S.C. § 2702 is the provision of the US Stored Communications Act that governs voluntary disclosure of communications content and customer records by electronic communication service providers. Subsection (b)(8) authorises a provider to disclose communications content to a governmental entity if the provider "in good faith, believes that an emergency involving danger of

trade-off: voluntary disclosure is permitted in narrow circumstances based on a good-faith assessment of an emergency, with reputational, legal, and political constraints discouraging both under-disclosure and over-disclosure. What is genuinely chatbot-specific is applying that good-faith standard to material that is more ambiguous and more intimate than the communications data the framework was originally designed for.

The second tension is between content-based safety monitoring and cryptographic hardening. Consider a chatbot embedded in WhatsApp, a messaging service whose central security promise to users is that the provider itself cannot read their messages end-to-end. If the embedded chatbot routes its queries to provider servers in plaintext, the embedded chatbot has, in effect, opened a window into a conversational environment that was specifically built to have no such windows. A standalone consumer chatbot operating outside any such environment inherits no analogous prior commitment, and the encryption question is an open product-design choice. The appropriate level of cryptographic protection is therefore not a single uniform answer for “chatbots” as a category; it depends on the prior architectural commitments that the deployment environment has already made to users, which the chatbot, by being embedded in or attached to that environment, inherits.

This second tension carries a further implication that is worth flagging directly. If a provider were to deploy cryptography that made the conversation genuinely inaccessible to the provider itself, content-based monitoring would become impossible, and provider-initiated disclosure would become a non-question: the provider cannot disclose what it cannot read, and the cases examined in Sections 1.2 and 1.4 would not have arisen in such an architecture, because the provider would have had no knowledge of the relevant conversations to disclose. The absence of knowledge would also significantly reduce the liability exposure that current providers face when monitoring fails or succeeds in contested ways. Safety, in that architecture, would come not from content monitoring but from upstream design: model-level filters and refusal behaviours that prevent the system from engaging with certain categories of content in the first place, before any conversation reaches the encryption layer. This is the longstanding “going dark” debate, but in an entirely new dimension. The classical debate concerned whether law enforcement could compel access to communications among third parties protected by encryption. The chatbot version concerns whether the provider itself ought to be in a position to read its own users’ conversations at all. The chapter does not propose to resolve that debate, but it is the frame within which the cryptographic recommendation below is best understood.

The third tension is between provider discretion in setting referral thresholds and the public’s ability to evaluate that threshold-setting work. Consider the Strasbourg case discussed in Section 1.2. A user’s conversation produced a real-time escalation through several intermediaries to French armed police, who intervened at the user’s home within hours. The question this chapter has identified is not whether the receiving authorities were accountable for what they did with the referral once they received it: that follow-on accountability operates through familiar mechanisms (warrants, civilian oversight, civil-rights litigation, prosecutorial review), and Section 1.7 sets this out in detail. The narrower question is whether the public is in a position to evaluate how the provider decided, before any of those downstream events, that this particular conversation warranted that particular escalation. Across the chatbot industry, that decision criterion is not currently publicly knowable. Providers operate under the good-faith requirement of § 2702, under reputational and political constraints, and within an information ecosystem that surfaces problematic cases over time. The narrower question this chapter identifies sits alongside those existing accountability mechanisms, not in opposition to them: it is whether aggregate transparency reporting (on the model that has developed in adjacent contexts, especially CSAM reporting) should be extended to provider-initiated emergency referrals in the chatbot setting, so that the criteria by which thresholds are set become available for public and scholarly assessment over time.

None of these tensions can be eliminated. All three can be governed more transparently than they currently are. The tentative recommendations that follow are addressed to providers, regulators, legislators, and the academic and practitioner communities that will need to do the substantive design

death or serious physical injury to any person requires disclosure without delay”; subsection (c)(4) authorises equivalent disclosure of customer records under the same standard. The good-faith requirement both protects the disclosing provider against liability (when the standard is met) and limits the lawful scope of disclosure (when it is not). There is no single EU-wide provision exactly parallel.

work; they are deliberately framed as directions for further work, given how fast the underlying practices are evolving and how recently the chatbot-specific issues have entered the public record.

Tentative Recommendations on Proactive Corporate Disclosure

Note: These recommendations address proactive corporate disclosure only. Recommendations on compelled government access and civil discovery are presented in their respective chapters. The Florida criminal investigation into whether ChatGPT bears criminal responsibility for the substance of its outputs to the FSU shooter, or the Tumbler Ridge case (Section 1.2 above) raise a further set of questions about provider liability that fall outside the scope of these recommendations.

#	Recommendation
1	Aggregate transparency reporting on provider-initiated emergency referrals. Major chatbot providers should consider introducing, into their existing transparency reports, aggregate data on the provider-initiated emergency referrals they make to law enforcement. The proper analogue is the proactive CSAM reporting these same providers already operate at scale. ⁷⁵ The reporting should include the volume of provider-initiated referrals, a categorical breakdown of the concerns that triggered them (violent threat to others, acute self-harm risk, other recognised emergency categories), and the jurisdictional distribution of the receiving authorities. The asymmetry between proactive reporting that already exists (for CSAM) and proactive reporting that does not (for violent threats and equivalent emergencies) is the specific gap this recommendation addresses.
2	Audit mechanisms for the referral pipeline, building on existing accountability structures. Existing accountability mechanisms (the good-faith requirement of § 2702 and equivalent EU provisions, civil-rights litigation, reputational and political constraints, the press, civil-litigation discovery, and the police-accountability framework that governs receiving authorities) are not absent from this landscape. The further question is whether the threshold-setting work that determines what enters the referral pipeline in the first place would benefit from periodic audit calibrated to the chatbot setting. Candidate models include audits of the criteria themselves (rather than of individual referral decisions, which raises difficult access and confidentiality issues); reviewable internal records of how the criteria are applied; and (where legally permissible) post-hoc user notification of provider-initiated emergency referrals. The design questions any such mechanism would raise (who audits, against what criteria, with what protections for the underlying material) warrant dedicated institutional work before a particular model is committed to.
3	Retention: distinguish preservation, user-elected storage, and provider-side retention. Three different concepts are sometimes conflated under the heading of retention. Preservation orders under established legal mechanisms ⁷⁶ operate within a well-established framework and are not the subject of this recommendation. User-elected storage (saved conversations, persistent memory features) is data the user has chosen to retain for their own access, and the principle of user control should govern its handling. Provider-side retention practices, by contrast, refer to provider-side logging beyond what the user has elected; they create a record that can serve both legitimate post-incident cooperation (as the FSU case

⁷⁵ The proper analogue is the proactive CSAM reporting that operates in a mandatory way through [NCMEC](#) under 18 U.S.C. § 2258A. Provider-initiated; aggregated; periodic; categorical breakdowns. Major chatbot providers already participate at scale (see Section 1.5 above). The transparency reports of Meta, X, and Google on emergency-disclosure requests received under § 2702(b)(8) measure a different and complementary activity (provider responses to law-enforcement-initiated requests) and illustrate that the methodological infrastructure for granular aggregate reporting is well established across the industry; see for example [Meta, Government Requests for User Data: Data Types](#). The recommendation extends the proactive-reporting practice (the CSAM model) into the violent-threat and equivalent emergency category.

⁷⁶ Preservation orders under 18 U.S.C. § 2703(f) and equivalent EU legal authorities operate within a well-established statutory framework and are not the subject of this recommendation; the recommendation addresses provider-side retention practices that are separate from formal preservation procedures triggered by specific legal process.

#	Recommendation
	illustrates) and purposes the user did not elect, and the tension between these two uses warrants further analytical work.
4	<p>Cryptographic hardening calibrated to the deployment environment. The level of cryptographic protection appropriate for a chatbot depends on the environment in which it operates, because that environment makes prior commitments to users that the chatbot inherits. Three deployment environments illustrate the point. A standalone consumer chatbot (ChatGPT in its own app) makes the security commitments customary for any online service (encryption in transit, access controls, encryption at rest on provider servers), but none of these prevent the provider itself from reading the user's conversations. Whether the system should additionally be designed to make conversations cryptographically inaccessible to the provider as well is, in that environment, an open product-design choice. A chatbot embedded in an end-to-end encrypted messaging service (Meta AI in WhatsApp being the leading example) operates inside an environment whose central security promise to users is that the provider cannot read their messages. Two architectural options arise: routing the embedded chatbot's queries to provider servers in a form readable to the provider, which opens a window into precisely the kind of conversations the surrounding environment was built to keep closed; or routing them through a confidential-computing pathway, which preserves coherence with the surrounding architecture. Meta AI in WhatsApp operated in the first mode for the year and a half following its introduction, and moved to the second on 13 May 2026 with the announcement of <i>Incognito Chat with Meta AI</i>.⁷⁷ The general architectural question, however, remains live for any future AI assistant embedded in an end-to-end encrypted environment. A chatbot operating under Sealed Mode (as proposed in Part 1 of this study) inherits the strengthened confidentiality commitment of that framework. The recommendation is that any policy discussion of cryptographic protection for chatbot conversations begin by identifying which of these architectural environments is in play, rather than treating all chatbot conversations as a single category. This position raises hard questions about how it would apply to specific products and deployment choices, and the chapter does not propose to settle the underlying encryption debate; the recommendation is that the deployment-environment dimension be made explicit in whatever debate does take place.</p>
5	<p>User-facing transparency about the nature of the chatbot relationship. Section 1.7 argued that the chatbot setting receives speech in the register of human confidant relationships but supplies almost none of the legal-institutional architecture (privilege, professional duties of care, social norms of confidentiality) that surrounds such speech in adjacent human contexts. This mismatch generates a duty owed by providers to users: to inform them, even more clearly than current practice does, of the conditions under which they are speaking, that conversations are recorded, may be subject to automated and human review, and may in defined circumstances be referred to law enforcement, and that the relationship does not carry the privileges, professional duties of care or confidentiality norms that surround the human relationships whose register it resembles. GDPR Articles 12-14 require some transparency disclosures, but these are not specifically calibrated to the chatbot context. Providers and regulators should develop chatbot-specific transparency disclosures that meaningfully communicate the operating conditions of the system at the point of use, in a form that is not lost among generic terms-of-service language.</p>

The five tentative recommendations above are deliberately limited in number and interdependent in substance. Taken individually, none on its own resolves the governance question this chapter has examined; taken together, they begin to set out a transparency-and-accountability architecture for

⁷⁷ For analysis of the *Incognito Chat with Meta AI* architecture and its implications for the deployment-environment typology, see Theodore Christakis and Peter Swire, "[The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI](#)", AI-Regulation Papers 2026-05-4, May 2026.

chatbot-context provider-initiated disclosure that builds on existing legal and institutional infrastructure rather than starting from scratch. Further work will be required to develop these recommendations into specific institutional designs.

2. When the State Compels Disclosure: Law Enforcement, Intelligence Agencies, and the Surveillance Multiplier

Chapter 1 examined what happens when chatbot providers decide, on their own initiative, to break the confidentiality of a user's conversation and refer it to law enforcement. This chapter examines what happens when the state does not wait for a voluntary referral. It compels disclosure.

The legal instruments are familiar: subpoenas, court orders, search warrants, preservation requests, intelligence directives. They have governed access to electronic communications for decades. But the material they now reach is not familiar at all. A government request served on a chatbot provider in 2026 can produce not merely a record of what someone said to another person, but a record of what someone thought, feared, planned, and felt, expressed in their own words, in a space many users experience as entirely private. When memory features are enabled, the request can produce a structured, continuously updated profile of the user's evolving inner life. When agentic capabilities are deployed, it can produce a log of actions the AI performed on the user's behalf across multiple external platforms. What the state reaches through a chatbot provider is therefore different in kind from the older categories of communications data: not a record of what a person communicated to others, nor a document or photograph they chose to store, but a record of their reasoning itself, drawn out in dialogue and set down in their own words in a setting they treat as private.⁷⁸

This chapter analyses the compelled-access landscape across five providers and six dimensions. Section 2.1 maps the basic legal ladder that governs routine government requests, from subscriber information to conversation content. Section 2.2 examines the procedural safeguards that determine whether the user is notified, whether the provider resists overbroad demands, and how cross-border requests are handled. Section 2.3 addresses preservation requests, which can silently freeze data that the user believes they have deleted, and shows how every retention-expanding feature documented in Part 1 is also a preservation-expanding feature, making retention not just a privacy issue but a surveillance multiplier. Section 2.4 discusses the important issue of reverse prompt warrants, in which the state searches not for a known suspect's data but for whoever entered a specific prompt. It documents the first known federal reverse prompt warrant served on an AI company, the rapidly fracturing constitutional case law on reverse keyword and geofence warrants, and the fact that the analytical convergence between reverse warrants and chatbot privacy is now being argued before the US Supreme Court. Section 2.5 turns to the foreign-intelligence dimension and the position of the hundreds of millions of non-US users, to whom the constitutional protections being litigated in the criminal cases do not extend. The natural concern is that FISA Section 702 might let intelligence agencies search chatbot prompt databases for content of interest; the chapter argues that this pathway in fact faces real statutory obstacles. Section 2.6 examines how memory and agentic AI transform the object of legal compulsion from a transcript into a structured personal dossier. Section 2.7 addresses the separation question raised by providers that simultaneously operate consumer chatbot platforms and provide AI technology to defence and intelligence agencies. Section 2.8 assesses transparency reporting and documents a bifurcated industry:

⁷⁸ None of this implies that other provider-held data is less intrusive. A full email archive or a cloud-stored photo library can be extraordinarily revealing, and compelling the disclosure of either raises serious privacy concerns of its own. The difference is one of character rather than degree. An email is addressed to a recipient and can be forwarded onward; a stored document or photograph is an artifact the user created and chose to keep; a chatbot conversation is neither, in that the user directs it at the system rather than to another person, so that, unlike an email, it reaches no further human addressee free to read it and pass it on, even though the provider holds it and can be compelled to produce it. It captures reasoning as it forms rather than a finished communication or record. A seized phone or laptop may well contain more than any chat log, including the chat logs themselves, but device seizure is a distinct legal mechanism from the compelled provider disclosure this chapter examines, and is reached by different means and usually with the user's knowledge.

dedicated AI-specific reporters (OpenAI, Anthropic) and providers whose chatbot data is statistically buried in broader corporate reporting or not reported at all.

Two developments give the chapter its focus. The first is the trajectory: government requests for chatbot data remain small in absolute terms beside legacy platforms, but they are climbing fast, with OpenAI moving from 6 in H1 2023 to 309 in H2 2025. The second, and the more important, is a change in kind. The early cases point to a shift from account-targeted requests, in which investigators seek a known user's data, toward content-based searches of provider databases, in which they ask who entered a given prompt. The technique is so far narrow and rare, but its logic, if extended, would turn the chatbot provider from a custodian of individual records into something closer to a searchable index of its users' conversations.

Before starting my analysis, a note on scope and method. The legal analysis in this chapter is grounded primarily in US law and procedure, for two reasons. First, four of the five providers examined are US-based, and the fifth (DeepSeek) is included not because Chinese law is comparably analysed here, but because excluding the most widely used non-Western chatbot from a comparative survey of consumer chatbot transparency would itself be analytically incomplete. Second, the most developed body of public case law, transparency reporting, and academic analysis of compelled access to chatbot data is US case law. The author is a European lawyer and does not purport to provide an authoritative account of Chinese civil procedure, Chinese national security law, or the relationship between DeepSeek and PRC authorities beyond what its own privacy policy publicly states. The treatment of DeepSeek throughout this chapter should accordingly be read as documenting publicly observable transparency posture, not as a substantive legal analysis comparable to the US discussion. Detailed analysis of the Chinese legal framework governing access to chatbot data is a task for scholars with the requisite expertise in Chinese law and would warrant a dedicated study.

2.1. The basic legal ladder

When providers do not disclose voluntarily, the state can compel them. The legal instruments are familiar: subpoena, court order, search warrant. What is new is the nature of the material they reach. Across the public policies of OpenAI, Anthropic, and Google, the common pattern follows the tiered structure that US law sets for electronic communications: basic subscriber information may be obtained with a subpoena or comparable legal instrument; non-content records may require a court order in some jurisdictions; and the content of communications requires a warrant or equivalent judicial authorisation.

OpenAI's January 2026 Government User Data Request Policy⁷⁹ is especially clear. It states that OpenAI discloses data only pursuant to valid legal process consistent with international human-rights laws, principles, and norms, or where it has a good-faith belief that there is an emergency involving death or serious physical injury. For OpenAI US, non-content information requires at least a subpoena, court order, search warrant, or equivalent, while user content is disclosed only in response to a valid warrant or equivalent. OpenAI also imposes procedural requirements that, while standard in US law, are worth noting because neither xAI nor DeepSeek publicly commits to equivalent specificity: particularity, account identification, and a defined time period. Notably, the threshold differs for European users: OpenAI Ireland, which serves users in the EEA and Switzerland, requires a judicially authorised order issued pursuant to Irish law before disclosing any user data, whether content or non-content, a higher standard than the subpoena threshold applied by OpenAI US for non-content information.

Anthropic's government-request reports use the same basic distinction, defining "content requests" as legal demands for content data such as prompts submitted to or responses provided by Anthropic's services, and "non-content requests" as demands for information such as name, email address, or other basic account information.⁸⁰ Google's government-request page similarly states that, under U.S. law,

⁷⁹ OpenAI, [OpenAI Government User Data Request Policy](#) (effective 1 January 2026).

⁸⁰ Anthropic, [Government Requests Report](#) (January–June 2025).

authorities must at least use a subpoena for basic subscriber information, a court order for non-content records in criminal cases, and a search warrant for the content of communications.⁸¹

For DeepSeek and xAI, the public picture is materially thinner. xAI's privacy policy says it may disclose personal information to comply with laws or respond to lawful requests and legal process.⁸² DeepSeek's privacy policy states that it may access, preserve, and share personal data with law-enforcement agencies or public authorities where it has a good-faith belief that this is necessary to comply with law, legal process, or government requests.⁸³ Neither company provides a public law-enforcement disclosure policy with the granularity offered by OpenAI or Google. The contrast is interesting: users of ChatGPT or Gemini can identify, before they type a single prompt, the legal threshold that the state must meet to obtain their conversations. Users of Grok or DeepSeek cannot.

Two recent US cases are a reminder that, even as this provider-facing ladder takes shape, chatbot content often reaches investigators by a quite different route. In California, federal prosecutors charged Jonathan Rinderknecht with igniting the New Year's Day 2025 fire that smouldered underground and days later rekindled into the Palisades fire, which killed twelve people and destroyed more than 6,800 buildings.⁸⁴ Among the evidence the US Attorney's office described was a dystopian image the defendant had generated on ChatGPT months earlier, depicting a burning forest with a crowd fleeing it, alongside questions he later put to the chatbot, including whether one is at fault if a fire starts from one's cigarette. That material was recovered from his own digital devices, not compelled from OpenAI. In Missouri, a Missouri State University student charged in early October 2025 with vandalising seventeen cars in a campus car park had, some ten minutes after the spree, confessed to ChatGPT and asked whether there was any way he could be identified; that conversation, too, was recovered from his phone rather than from the provider.⁸⁵ In both, the chatbot material reached investigators through the user's own device, not through compelled disclosure from the provider, the distinct route noted at the outset of this chapter. What the cases establish is therefore not how the ladder operates but how much evidentiary weight chatbot conversations now carry: as Richard Salgado has observed, prosecutors increasingly treat AI chat content with the seriousness once reserved for email and other non-public communications.⁸⁶

2.2. User notice, resistance, and cross-border process

The second dimension of compelled access concerns procedural safeguards: does the provider notify the user? Does it resist defective or overbroad demands? Does it distinguish between jurisdictions? On each of these questions, the divergence among providers is significant.

OpenAI states that it provides notice to users before disclosure unless prohibited by law, and that it may withhold notice in exceptional circumstances such as emergencies or compromised accounts, while providing delayed notice when the prohibition expires.⁸⁷ Anthropic similarly emphasises three principles: careful legal review; notice to users unless legally prohibited or in extraordinary situations; and, most distinctively, a stated preference that government agencies seek data directly from the customer rather than through Anthropic, wherever possible.⁸⁸ Google states that it emails the user account before disclosing information, subject to legal prohibitions, emergencies, and delayed notice after gag periods expire.⁸⁹

⁸¹ Google, "[How Google handles government requests for user information](#)".

⁸² xAI, [Privacy Policy](#).

⁸³ DeepSeek, [Privacy Policy](#) (last updated 10 February 2026).

⁸⁴ On the Palisades fire prosecution, see CNN, "[How ChatGPT prompts and a lighter led Palisades Fire investigators to a suspected arsonist](#)", (16 October 2025), and Axios, "[Palisades arson suspect allegedly used ChatGPT to create images of burning city](#)" (8 October 2025); the US Attorney for the Central District of California stated that the ChatGPT image and related queries were among the evidence collected from the defendant's digital devices.

⁸⁵ On the Springfield case, see Ozarks First, "[ChatGPT, cell data help arrest Springfield teen for MSU parking lot vandalism](#)" (1 October 2025), and The Register, "[Student admits vandalism spree to ChatGPT, cops say](#)" (2 October 2025); the probable-cause statement records that the ChatGPT conversation was recovered from the defendant's phone.

⁸⁶ See Matt Fleischer-Black, "[Gen AI Chats Becoming Evidence: Law Enforcement Warrants and Subpoenas](#)", *Cybersecurity Law Report*, 3 December 2025.

⁸⁷ OpenAI, Government User Data Request Policy, *op. cit.*, Section V.

⁸⁸ Anthropic Report, *op. cit.*

⁸⁹ Google policy, *op. cit.*

Google and Anthropic also expressly say that they narrow, challenge, or reject overbroad or legally deficient requests. OpenAI's policy requires particularity and limits production to the data specified in the legal process.

By contrast, xAI's public materials do not identify a comparable consumer-facing user-notice policy for Grok. An additional complexity arises from xAI's dual-access structure: users who access Grok through the X platform are governed by X's privacy policy, not xAI's. X maintains a transparency centre with reporting on government requests, but this covers the X/Periscope ecosystem and does not separately identify Grok-related demands. Users who access Grok through grok.com or the Grok mobile app are governed by xAI's privacy policy, which contains no equivalent transparency reporting. The same chatbot thus operates under different disclosure regimes depending on how the user reaches it.

DeepSeek's privacy policy contains no comparable advance-notice or resistance commitments.

The cross-border dimension is analytically significant because it determines which legal system governs a request and whether mutual legal assistance or other government-to-government mechanisms are required. OpenAI expressly distinguishes between OpenAI US and OpenAI Ireland, with EEA and Swiss users served by OpenAI Ireland under Irish law. Requests from authorities outside the relevant jurisdiction must often proceed through MLATs, executive agreements satisfying the CLOUD Act, letters rogatory, or other government-to-government channels, though OpenAI leaves room in some cases for direct recognition of foreign legal process consistent with domestic law and international human-rights norms. Google follows a similar dual-entity logic. Anthropic's cross-border framework is less publicly detailed: its government-request reports articulate principles of careful legal review and user notice but do not describe a jurisdictional architecture comparable to OpenAI's US/Ireland structure or Google's dual-entity model. DeepSeek presents the sharpest contrast: its privacy policy states that it directly collects, processes, and stores personal data in the People's Republic of China, which materially affects the legal and strategic landscape for every user outside China.

Table 2. Comparative overview of compelled government access

Dimension	ChatGPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	Grok (xAI)	DeepSeek
Public legal-process framework	Detailed dedicated law-enforcement policy (January 2026). Distinguishes OpenAI US / OpenAI Ireland.	Detailed periodic reports plus articulated principles (careful review, user notice, preference for direct customer access).	Detailed Google-wide legal-request policy with long track record.	No equivalent dedicated Grok law-enforcement policy located in public materials.	No dedicated law-enforcement request policy; governed mainly by privacy policy. Data stored in PRC.
Content vs. non-content distinction	Explicit. Content requires warrant or equivalent; non-content requires subpoena/order/warrant depending on category.	Explicit. Reports distinguish content, non-content, and emergency requests.	Explicit in Google-wide policy (subpoena / court order / warrant ladder).	Not publicly spelled out with comparable granularity.	Not publicly spelled out with comparable granularity.
User notice	Yes, before disclosure unless prohibited by law; delayed notice where possible after gag expiry.	Yes, unless legally prohibited or in extraordinary emergencies involving imminent risk.	Yes, before disclosure, subject to gag orders, emergencies, and disabled/hijacked accounts.	No comparable Grok-specific advance-notice commitment located.	No comparable advance-notice commitment identified.

Stance toward overbroad requests	Requires particularity, account identification, and defined time period. Produces only specified data.	Publicly states it challenges or rejects overbroad, vague, or legally deficient requests.	Publicly states it narrows or objects to excessive requests.	Public materials do not provide equivalent detail.	Public materials do not provide equivalent detail.
Cross-border structure	OpenAI US / OpenAI Ireland split. MLAT, CLOUD Act, or government-to-government channels often required for cross-border requests.	Applicable-law framing, but less publicly detailed jurisdictional architecture than OpenAI or Google.	Google LLC / Google Ireland service-provider structure. Foreign requests must satisfy US law and local due-process requirements.	US-based xAI service; public cross-border request framework not described in comparable detail.	Data stored in PRC. Subject to Chinese national security and intelligence cooperation laws. Cross-border framework unclear.
Preservation requests	Honours valid 18 U.S.C. § 2703(f) requests. 90-day preservation, extendable for additional 90 days.	Not publicly detailed with equivalent specificity.	Google-wide preservation framework applies.	Privacy policy allows retention for legal reasons including court orders.	Retains data as long as necessary for stated purposes including legal obligations.
National security process	FISA/NSL reported in broad bands only: 0–249 requests, 0–249 accounts (H2 2025).	NSL/orders/directives reported in bands with one-year delay: 0–99 requests, 0–99 accounts (2024).	Reported within Google-wide national security bands.	Reported within X-wide national security bands.	Subject to Chinese national security and intelligence cooperation laws. No public reporting.

Note: Assessments based on publicly available materials as of April 2026. Policies may have changed since review.

2.3. Preservation and the practical defeat of ephemerality

Compelled access frequently begins not with a disclosure order but with a preservation request. Under 18 U.S.C. § 2703(f), a governmental entity can require an electronic communication service provider to preserve records pending the issuance of a court order or other process. OpenAI’s policy confirms that it honours such requests, preserving user data for up to 90 days and extending the preservation for an additional 90 days upon a renewed request. These are statutory time limits designed to give the government a window in which to obtain formal legal process. A preservation request is not itself an authority to compel disclosure; it freezes the data in place so that a subsequent subpoena, court order or warrant, of the kind set out above, can reach it. Critically, the user is not notified that their data has been frozen. A user who deletes a conversation during the preservation period may see the conversation disappear from their interface, but the provider is legally obligated to retain the data on the backend regardless. For the duration of the preservation obligation, deletion is no longer deletion in the ordinary user-facing sense: the record survives on the backend even if it disappears from the interface.

This has two consequences that bear directly on the privacy analysis in Part 1. First, the “deletion” and “limited retention” features that providers market to consumers, including temporary chat modes and 30-day retention windows, can be overridden by a government request. Second, the more data a provider stores at the moment a preservation request arrives, the more data the state can freeze. Memory features, conversation histories, uploaded files, connected health records and agentic action logs all

enlarge the pool of data available to be frozen. The relationship is simple, and not automatic: a preservation request reaches only what the provider happens to hold when it arrives, but the more a provider retains, the more a single request can sweep in. Each retention-expanding feature documented in Part 1 therefore enlarges what a preservation can capture, and every preservation is a precursor to potential compelled disclosure. As consumer chatbots become more persistent and agentic, the volume of sensitive material exposed to that mechanism rises with them.

A civil case makes the same point from a different direction, and at a scale no §2703(f) request approaches. In the New York Times copyright litigation against OpenAI, analysed in Chapter 3, a magistrate judge ordered the company in May 2025 to preserve and segregate output-log data that would otherwise have been deleted, including conversations users had deleted or set as temporary; the district judge affirmed the order on appeal, and OpenAI objected publicly on privacy grounds. OpenAI's obligation to preserve on a going-forward basis ended on 26 September 2025, after which it returned to its standard practice of deleting conversations within thirty days, though the logs already preserved remain within the plaintiffs' discovery set. Then, in December 2025, the court went further and ordered OpenAI to produce a sample of twenty million de-identified consumer conversation logs in discovery. The mechanism here is civil discovery, not the criminal process or the §2703(f) preservation that is this section's subject, and the rules differ accordingly. But the episode shows, concretely and at scale, that legal process can override the deletion and temporary-chat features users rely on, and can do so without notifying the affected users, whose attempts to intervene were denied because they were not parties to the case.

2.4. Reverse prompt warrants: when the state searches for the user, not the data

Traditional government access to electronic communications follows a familiar model: investigators identify a suspect, then seek that suspect's data from the provider. Reverse prompt warrants invert this logic. Instead of targeting a known individual, the government asks the provider to search its database for any user who entered a specific prompt or sequence of prompts, and to produce the identifying information of whoever matches.

The first known federal warrant compelling an AI company to identify a user from their prompts was reported by Forbes in October 2025,⁹⁰ which, having reviewed the unsealed court records, described it as the first such warrant served on OpenAI; Riana Pfefferkorn at Stanford's Center for Internet and Society analysed it in detail.⁹¹ The warrant, issued in Maine in a child exploitation investigation, sought to identify a ChatGPT user by matching two specific prompts that the suspect had shared during undercover communications. The two prompts led to a single account. OpenAI apparently complied by providing an Excel spreadsheet of user data. In this instance the warrant was narrowly tailored, particularised, and linked to a pre-existing investigation. The constitutional concerns lie in what comes next.

Two scenarios should be distinguished, because the constitutional questions they raise are not the same. The technique is identical in both: the provider is asked to search its prompt database for content supplied by an unknown person and to return whoever matches. What differs is the breadth of the search and whether it is anchored to a pre-existing suspect. The first, illustrated by the Maine warrant, is the narrowly drawn reverse prompt warrant: the prompts sought are specific enough to identify a single user, and the warrant works in effect as a content-based way to confirm or rule out an existing investigative lead. That the first known warrant appears to have returned a single account does not lessen the novelty of the method, since the user was still found by searching a provider's database for content rather than by being named in advance. The second, so far hypothetical, is the genuinely bulk reverse prompt warrant: a demand that the provider search its entire user base for anyone whose conversations match a content pattern, a topic or theme for instance, with no prior link to a particular suspect. The pre-existing nexus is the hinge. Its presence in the first scenario supplies the individualised

⁹⁰ "[DHS Ordered OpenAI To Share User Data In First Known Warrant For ChatGPT Prompts](#)", Forbes (Oct. 20, 2025).

⁹¹ Riana Pfefferkorn, "[Eight \(or so\) Questions to Ask About the ChatGPT Warrant](#)", Stanford CIS Blog, 24 October 2025.

suspicion and particularity the Fourth Amendment requires; its absence in the second is what makes the concerns about particularity, probable cause and overbreadth escalate sharply.

As Pfefferkorn has observed, the critical question is whether future reverse prompt warrants will be as narrow as the first one, or whether prosecutors will “judge-shop until they get a friendly judge who’ll sign off on one that isn’t narrowed down to one specific user”.

The analogy to “reverse keyword warrants” (which ask Google to identify everyone who searched for a specific term) and “geofence warrants” (which ask for all users whose devices were in a specific location) is direct,⁹² but the material a reverse prompt search returns is typically richer. A web search is usually a brief query entered to retrieve documents; a chatbot session is a conversation, carried on in sentences and over many turns, in which users tend to elaborate, reason aloud and confide. The *amici* in Chattrie drew the same contrast, noting that chatbot prompts can be “far more detailed and revealing” than search queries and are designed to sustain a dialogue touching a user’s most intimate thoughts.⁹³ The self-disclosure research discussed in the Introduction points the same way. This is a difference of degree rather than a clean line: a sequence of search queries can itself expose planning and state of mind, which is precisely why reverse keyword warrants are contested. But the difference is real, and it means a reverse prompt search can convert mere curiosity, fictional drafting or poorly contextualised research into an investigative lead more readily than a single matched search term.

In a subsequent March 2026 lecture, Pfefferkorn and other scholars explored how federal digital communications privacy law (the Stored Communications Act, or SCA, which is part of ECPA) applies to AI providers.⁹⁴ Paul Ohm has argued that reverse searches are likely not authorised under the SCA and may also violate the Fourth Amendment.⁹⁵

US courts are actively and divisively grappling with the legality of reverse searches, and the outcomes will directly shape the legal environment in which reverse prompt warrants operate.

On reverse keyword warrants, the first supreme court to address the question was the Colorado Supreme Court in *People v. Seymour* (2023).⁹⁶ The court found that the defendant had a constitutionally protected privacy interest in his Google search history, even when revealed only in connection with his IP address and not his name, and held that reverse keyword searches implicate freedom of expression, requiring constitutional protections to be applied with “scrupulous exactitude”. The court assumed without deciding that the warrant was constitutionally defective for lack of individualised probable cause, but admitted the evidence under the good-faith exception, reasoning that law enforcement had acted reasonably in deploying a novel technique whose legality had not yet been adjudicated. In December 2025, the Pennsylvania Supreme Court reached a sharply different result in *Commonwealth v. Kurtz*. In a fractured decision, an opinion announcing the judgment, joined by three justices, concluded that the defendant had no enforceable expectation of privacy in his Google search terms, reasoning that by entering a query into Google he had **voluntarily conveyed the contents of the search to a third party**. Three justices concurred in the result on probable-cause grounds, assuming without deciding that a constitutionally protected search had occurred, while one justice dissented and would have recognised a reasonable expectation of privacy under the Pennsylvania Constitution.⁹⁷ The case law on keyword warrants is therefore actively splitting: in some jurisdictions, reverse searches of chatbot prompt databases could face serious Fourth Amendment challenges; in others, the third-party doctrine could be held to extinguish the user’s privacy interest entirely.

⁹² For a comprehensive overview of the constitutional landscape governing both geofence and keyword warrants, see Congressional Research Service, “[Geofence and Keyword Searches: Reverse Warrants and the Fourth Amendment](#)”, R48852 (February 2026).

⁹³ See the Brief of the Center for Democracy & Technology et al. (*infra* note 103): “A corollary to reverse keyword searches with potentially even more profound dangers are reverse warrants for AI chatbot prompts. [...] While chatbots perform a similar role to search engines in providing responsive information to a query, users’ prompts can be far more detailed and revealing, and designed to facilitate an ongoing “dialogue,” potentially regarding the user’s most intimate thoughts and ideas”.

⁹⁴ Riana Pfefferkorn, “[Applying Federal Digital Communications Privacy Law to Hosted AI Models](#)”, Stanford CIS Blog, March 2026.

⁹⁵ Paul Ohm, discussing reverse searches and AI chatbots as a “gold mine of evidence” in the context of the Stored Communications Act. Cited in Pfefferkorn, Stanford CIS Blog, March 2026.

⁹⁶ *People v. Seymour*, 536 P.3d 1260 (Colo. 2023). See EFF, ‘Colorado Supreme Court Upholds Keyword Search Warrant,’ 18 October 2023; State Court Report, ‘Colorado Supreme Court Upholds Controversial Google Keyword Warrant,’ 20 October 2023. See also NBC News, “[Police used a reverse keyword Google search to find an accused killer. He says that’s illegal](#)”, 1 July 2022.

⁹⁷ *Commonwealth v. Kurtz*, No. 98 MAP 2023 (Pa. Dec. 16, 2025). Emphasis added. See National Law Review, “[Can Law Enforcement Access Google Search Data Without a Warrant? Pennsylvania Says Yes](#)”, December 2025.

The third-party doctrine on which that second outcome rests is not the fixed rule it once was. Established in *Smith v. Maryland* for the numbers dialed from a telephone and in *United States v. Miller* for bank records, it holds that information voluntarily handed to a third party carries no reasonable expectation of privacy against the government.⁹⁸ In *Carpenter v. United States* (2018), however, the Supreme Court declined to extend the doctrine mechanically to historical cell-site location data, reasoning that records which are deeply revealing, comprehensive, and generated without any meaningful act of choice fall outside its logic.⁹⁹ Chatbot prompts arguably satisfy those conditions at least as strongly as the location data in *Carpenter*: a memory-backed conversational record is more revealing than a map of a person's movements, and providers retain it at scale. The point on which the argument turns is the one *Carpenter* left most exposed, voluntariness, because a user affirmatively composes a prompt, which can look closer to the cheque handed to a bank in *Miller* than to the location signal a phone emits automatically. This is where architecture matters: a conversation conducted under a Sealed Mode that verifiably constrains the provider's own access has not, in any meaningful sense, been voluntarily conveyed to the provider as a third party for its use, which weakens the premise on which *Smith* and *Miller* rest and brings the case within *Carpenter*'s logic. *Carpenter* was an expressly narrow decision, so this remains an argument to be made rather than a settled result; but it is the strongest constitutional foothold available, and it is strongest for exactly the sealed, high-sensitivity conversations this study argues should be protected. *Carpenter* was itself a location case, which is why the question recurs with full force in the geofence litigation now before the Court.

On geofence warrants, the question has now reached the US Supreme Court. In *Chatrue v. United States*, the Court heard oral arguments on 27 April 2026 on the constitutionality of a geofence warrant that required Google to search its Location History database in order to identify users whose devices had been within a specified area near a bank robbery, and ultimately to produce identifying information for a selected subset of those users.¹⁰⁰ The lower courts have produced sharply divergent results. The district court held that the warrant violated the Fourth Amendment but declined to suppress the evidence under the good-faith exception. The Fourth Circuit, sitting en banc, affirmed the district court's denial of suppression in a one-sentence per curiam opinion, with the en banc court dividing seven-to-seven on whether a Fourth Amendment search had occurred and producing nine separate concurring and dissenting opinions across more than one hundred pages. The Fifth Circuit, in a separate case, held that geofence warrants are "modern-day general warrants" that violate the Fourth Amendment categorically.¹⁰¹ The case law on geofence warrants is therefore deeply unsettled, and the Supreme Court's eventual decision will shape the legal environment in which the closely related question of reverse prompt warrants is decided.¹⁰²

What matters most here is not the outcome of *Chatrue* but what the case reveals about where the law is heading. The Brennan Center for Justice, warned that upholding the geofence warrant would enable a new class of reverse search warrants, including warrants that could identify **"everyone who... issued a prompt to an AI chatbot"**.¹⁰³ Geofence warrants, reverse keyword warrants and reverse prompt

⁹⁸ *Smith v. Maryland*, 442 U.S. 735 (1979); *United States v. Miller*, 425 U.S. 435 (1976).

⁹⁹ *Carpenter v. United States*, 585 U.S. 296 (2018). The Court emphasised that its holding was a narrow one and did not disturb the application of the third-party doctrine to conventional business records.

¹⁰⁰ *Chatrue v. United States*, No. 25-112 (U.S., oral argument heard 27 April 2026). Decision expected by summer 2026. On Google's architectural response to geofence warrants (migrating location history data from centralised servers to users' individual devices to make geofence warrant compliance technically impossible), see TIME, "[Google Will Stop Providing Law Enforcement Data on Which Users Were Near a Crime](#)", 15 December 2023.

¹⁰¹ *United States v. Smith*, 110 F.4th 817 (5th Cir. 2024), cert. denied, 146 S. Ct. 356 (2025). On the procedural history of *Chatrue* in the lower courts, see Brookings Institution, "[Supreme Court weighs constitutionality of geofence warrants](#)", 29 April 2026.

¹⁰² SCOTUSblog, "[Justices appear mixed on whether geofence warrant violated the Fourth Amendment](#)", 27 April 2026; Orin Kerr, "[A Few Thoughts on the Chatrue Oral Argument](#)", Volokh Conspiracy (Reason), 27 April 2026. Early commentary following oral argument suggests that the Court may avoid a categorical rule either way, instead focusing on scope, particularity, and how much judicial supervision is required for reverse-location searches.

¹⁰³ Brennan Center for Justice, [Okello Chatrue v. United States of America](#) (18 March 2026) (explaining that the coalition's amicus brief warns that authorising the geofence warrant would enable a new class of reverse-search warrants that "may enable dragnet searches for everyone who... issued a prompt to an AI chatbot"). The brief is the [Brief of the Center for Democracy & Technology, the Surveillance Technology Oversight Project, the Brennan Center for Justice and Defending Rights & Dissent as Amici Curiae in Support of Petitioner](#), [*Chatrue v. United States*, No. 25-112] 27-30 (arguing that the geofence warrant lacked particularity and probable cause, and that upholding it would facilitate a broader class of reverse warrants, including reverse warrants for AI chatbot prompts, "a corollary to reverse keyword searches with potentially even more profound dangers"). See also [Brief of Eight Law Professors as Amici Curiae in Support of Petitioner](#), [*Chatrue v. United States*, No. 25-112], Part II.C ("AI Chatbots") (extending the same reverse-warrant analysis to AI chatbot prompts, and citing both the Rinderknecht complaint and *United States v. Heppner*).

warrants are variants of one technique: compelling a provider to search its whole database to find the unknown people who match a pattern the police specify. The constitutional rules the Court settles for one will therefore largely govern the others, which is why the fate of chatbot prompts is now bound up with a bank-robbery geofence case before the highest court in the United States.

Those Fourth Amendment protections, however, generally do not extend directly to non-US persons located outside the United States. Under *United States v. Verdugo-Urquidez*, 494 U.S. 259 (1990), the Fourth Amendment applies only to persons who have developed “substantial connections” with the United States, a test that excludes most ordinary users abroad.¹⁰⁴ In the criminal investigation context, the constitutional doctrine being developed in *Chatrie*, *Seymour* and *Kurtz* nevertheless seems to do indirect work on behalf of all users in a provider’s database: the Stored Communications Act requires a warrant for disclosure of communications content, and that warrant, like any federal warrant, must satisfy Fourth Amendment particularity requirements regardless of whose data the database contains. If those cases ultimately establish that reverse prompt warrants lack particularity, warrants of that kind would be constitutionally defective; courts would not issue them, and providers would have grounds to challenge any that were nevertheless issued. The practical effect would be to protect all users in the database, regardless of nationality. If the cases instead uphold reverse prompt warrants, federal criminal process becomes a pathway to compelled reverse searches of chatbot databases for any user, US or non-US.

There is, however, a response that does not wait on how those cases are resolved. Google’s own answer to the geofence-warrant problem was architectural rather than legal: in December 2023 it moved Location History to encrypted, on-device storage by default, so that it no longer holds the central database such a warrant would search and can no longer comply with one.¹⁰⁵ Where the provider cannot read or search the corpus, the reverse-search question is answered by design, and for every user in the database regardless of nationality. The same architectural logic has now reached the chatbot setting: Meta’s Incognito Chat with Meta AI, launched on 13 May 2026, routes conversations inside WhatsApp through a Trusted Execution Environment the provider cannot read, so that those exchanges too lie beyond the reach of any reverse search.

The position of non-US users under US law raises a separate question in a different setting: foreign intelligence collection, where the warrant procedure does not apply at all and the protections are statutory and executive rather than constitutional. The natural worry is that Section 702 of FISA might supply what the criminal warrant cannot. The next section examines that worry directly, and finds it more limited than it first appears.

2.5. The foreign intelligence dimension: Could Section 702 Be Used to Compel Reverse-Prompt Searches?

The previous section concerned criminal process, and the protections it examined run, for the most part, only to people with substantial connections to the United States. That leaves an obvious question for the hundreds of millions of users outside the country who rely on US-based chatbots: what does foreign-intelligence law expose them to? The instrument to enquire about is Section 702 of the Foreign Intelligence Surveillance Act, the authority under which US agencies compel providers to help acquire the communications of non-US persons abroad, and which now runs through a large share of routine intelligence work.¹⁰⁶ If criminal warrants might one day be read to allow reverse searches of chatbot databases, could Section 702 allow the same against foreign users, and without any warrant at all?

¹⁰⁴ *United States v. Verdugo-Urquidez*, 494 U.S. 259, 265 (1990), holding nonetheless that the Fourth Amendment’s protection of “the people” extends only to persons who have developed “substantial connections” with the United States; the search of a non-resident alien’s property in a foreign country by US agents was therefore not subject to the Fourth Amendment. The “substantial connections” formulation has been understood, especially through Justice Kennedy’s concurrence, as the operative test in subsequent cases.

¹⁰⁵ On Google’s decision to store Location History on users’ devices in encrypted form, so that it could no longer respond to geofence warrants, see Cyrus Farivar and Thomas Brewster, “[Google Just Killed Warrants That Give Police Access to Location Data](#)”, *Forbes* (14 December 2023); Jennifer Lynch, “[Is This the End of Geofence Warrants?](#)”, Electronic Frontier Foundation (13 December 2023); and *TIME*, “[Google Will Stop Providing Law Enforcement Data on Which Users Were Near a Crime](#)” (15 December 2023).

¹⁰⁶ The PCLOB’s 2026 staff report notes that 63% of the articles in the President’s Daily Brief in 2025 contained Section 702 information, up from 59% in 2024. Privacy and Civil Liberties Oversight Board (PCLOB), [Report on Section 702 \(2026\)](#), at 4.

The short answer, set out below, is that Section 702 appears poorly suited to this particular task. That is not a claim that non-US users are well protected in general, nor that the authority could never be stretched by argument or future amendment; it is a narrower point about one specific operation, the generalised content-based reverse search, and why the statute as written does not sit easily with it. This is contested terrain, and what follows is offered as a reasoned reading rather than a settled answer.

Section 702 rests on a different bargain from the criminal warrant regime.¹⁰⁷ Under a criminal warrant the gate is high, probable cause, individualised judicial review and particularity, but the operational scope behind that gate is broad. Under Section 702 the gate is, for non-US persons abroad, essentially absent: no individualised judicial review of targeting, no probable cause, no warrant. What the statute puts in place of that missing gate is a set of limits on what may be done, and in particular a requirement that acquisition proceed through selectors tied to specific targets. It is those limits, not any warrant requirement, that bear on the question here.

¹⁰⁷ Section 702 of the Foreign Intelligence Surveillance Act (50 U.S.C. § 1881a), as reauthorised by the Reforming Intelligence and Securing America Act (RISAA), 20 April 2024. See Congressional Research Service, [FISA Section 702 and the 2024 Reforming Intelligence and Securing America Act](#), R48592 (8 July 2025).

Could Section 702 compel a generalised reverse-prompt search?

The operation worth testing is the most expansive one: a demand that a chatbot provider search its prompt database for any non-US user whose conversations contain specified content, a particular topic, technique or name of interest, and hand over whoever matches.¹⁰⁸

On the statute as written, this runs into real obstacles. Section 702 authorises the targeting of non-US persons reasonably believed to be abroad through selectors associated with those targets, and the compelled-assistance authority of Section 702(i) reaches the acquisition of *those* targets' communications, not open-ended content searches across a provider's user base.¹⁰⁹ A content-based sweep of prompts has no target and no selector; it starts from the content itself. That is the feature that places it outside the natural reading of the authority.

The reading is reinforced by RISAA's treatment of 'abouts' collection, the former practice of intercepting communications whose content contained a tasked selector rather than communications to or from one. RISAA bars any resumption of 'abouts' collection, with no exceptions.¹¹⁰ A content-based reverse search of chatbot prompts resembles 'abouts' collection in the respect that matters: it picks out communications by what they contain, not by their link to an identified target. Whether a court or the Foreign Intelligence Surveillance Court would treat that resemblance as legally decisive is a question for specialists in the regime, and not one I can resolve here. But it points the same way as the selector requirement: Section 702 is not naturally an authority for generalised content searches across a provider's entire user base. Should that reading prove too confident, other limits would still apply to any such collection.¹¹¹

A note on the apparent asymmetry

If that reading is right, it produces a result worth pausing over. Should the constitutional cases in Section 2.4 eventually be read to permit not only narrowly drawn reverse-prompt warrants but bulk content-based ones, criminal process would supply a route to compelled reverse searches of chatbot databases that Section 702, for the reasons above, seems poorly suited to authorise. The regime usually thought the more permissive, intelligence collection against foreigners, would on this one operation be the more constrained.

The asymmetry is real but explainable. The two regimes carry different kinds of authority, hedged by different kinds of limits. The criminal regime allows broad scope behind a demanding gate; Section 702 allows collection against non-US persons abroad with no individualised gate, but pays for that absence with limits on the kinds of operation it can compel. A pure reverse search being hard to fit within Section 702 is that bargain doing its intended work. None of this means non-US users are protected across the board. They sit outside most of the constitutional protections litigated in Section 2.4, and the wider

¹⁰⁸ The intelligence interest could be illustrated by these entirely hypothetical examples: an agency might direct a chatbot provider to identify all users who have submitted prompts discussing specific weapons systems, nuclear enrichment techniques or missile guidance architectures; it might search for prompts discussing sanctions evasion, cryptocurrency laundering methods or means of transferring funds to designated entities; it might search for prompts mentioning particular individuals, organisations or locations of intelligence interest; or it might target foreign users who have discussed strategies for circumventing US tariffs or export controls, or who have used a chatbot to draft communications relating to trade negotiations or industrial policy in sectors the US considers strategically significant.

¹⁰⁹ Section 702 is not free of judicial involvement; that involvement operates at the level of the programme rather than the individual target. Each year the Attorney General and the Director of National Intelligence submit certifications to the Foreign Intelligence Surveillance Court, which reviews and must approve the accompanying targeting, minimization and querying procedures before collection may proceed, and compliance with those procedures is then assessed every six months and reported to the Court and to the congressional intelligence and judiciary committees. What is absent, by contrast with the criminal warrant regime, is *ex ante* individualised judicial review of particular targeting decisions. See Congressional Research Service, [FISA Section 702 and the 2024 Reforming Intelligence and Securing America Act](#), R48592 (8 July 2025).

¹¹⁰ Reforming Intelligence and Securing America Act, Pub. L. No. 118-49, 138 Stat. 882 (20 April 2024). Under RISAA, the government is barred from resuming 'abouts' collection under Section 702, with no exceptions. See Congressional Research Service, [FISA Section 702 and the 2024 Reforming Intelligence and Securing America Act](#), R48592 (8 July 2025), describing how RISAA codified the prior NSA decision to discontinue 'abouts' collection and removed the administrative discretion to resume it, and noting that 'under the RISAA, the government is barred from resuming abouts collections under Section 702, with no exceptions'.

¹¹¹ Even if the analysis above were thought to overstate the obstacles to compelled reverse searches under Section 702, other safeguards would continue to apply. Executive Order 14086, adopted in October 2022 as part of the legal architecture underpinning the EU-US Data Privacy Framework, imposes necessity, proportionality, purpose-limitation and prohibited-objectives constraints on signals intelligence collection, and provides a redress mechanism through the Civil Liberties Protection Officer and the Data Protection Review Court.

question of whether one regime can quietly become a route to what the other forbids is a real and recurring one in US surveillance law. The narrow conclusion here is only that, on this particular operation, the intelligence route appears the tighter of the two, a conclusion I hold with appropriate humility given how much of the governing practice remains classified.

2.6. Memory, agentic AI, and the intelligence “gold mine”

Part 1 of this study identified memory as the emerging fifth boundary of chatbot privacy. Several providers now offer persistent personalisation features that build longitudinal user profiles from accumulated chat interactions: preferences, health concerns, professional strategies, family circumstances, emotional patterns. As chatbots evolve into agentic systems that can browse the web, manage files, connect to external services, and execute tasks on the user’s behalf, the object of legal compulsion changes qualitatively.

A traditional government request targets a chat transcript: what the user said and what the model responded. A request targeting memory targets something far more valuable: a structured, provider-curated summary of who the user is, what they care about, what they plan, and how they think. If an intelligence agency were to gain access to these memory files, it would possess not a raw transcript but a synthesised, structured profile of a user’s evolving thoughts, preferences, and social connections.

The comparison to advertising profiles is instructive but insufficient. Social media platforms and data brokers construct profiles primarily from observed behaviour: clicks, purchases, location data. While such profiles also incorporate the public posts and profile information that users contribute, the directly-disclosed component is shaped by the user’s awareness of a public or semi-public audience and does not generally extend to private reflection. Chatbot memory profiles are different because they are built from direct, explicit, often intimate self-disclosure addressed to the system itself. They can combine what the user does with what the user says they fear, want, plan, believe, or need. That combination of behavioural inference and explicit private self-description in a single searchable profile is what makes chatbot memory unusually sensitive.¹¹²

Could intelligence agencies use FISA Section 702 to obtain the memory files of a specific non-US person reasonably believed to be abroad and validly targeted for foreign-intelligence purposes? This turns on a set of conditions, not on the absence of any bar. The provider would first have to qualify as an electronic communication service provider (ECSP), a category RISAA’s 2024 amendments widened but whose application to AI chatbot providers is unsettled; the usual Section 702 prerequisites would then apply (a foreign-intelligence purpose, a non-US target reasonably believed to be abroad, acquisition under the certified targeting and minimization procedures); and the user would have to have enabled memory, with the provider holding it in a form it can access and produce. Where those conditions are met, the statute as written does not obviously stand in the way: Section 702 authorises the acquisition of foreign intelligence information from the communications of targeted non-US persons abroad, and a directive could in principle require a qualifying provider to produce a target’s stored memory. Whether a provider-curated profile is itself a “communication” within that authority, rather than a derived record, is a further question the public materials do not resolve. Such a file can be more revealing than an equivalent volume of email or intercepted calls, because it is a structured synthesis of a user’s preferences, plans and patterns of thought drawn from many exchanges; in other cases it will be less so, where the intercepted content is itself rich and the memory profile thin. The point is not that the memory file is always the more revealing record, but that it is a different kind of record, curated and condensed, and that this is what such a demand would reach.

¹¹² One architectural response to the surveillance concerns developed in this section is to renounce persistent memory across sessions altogether. Meta’s *Incognito Chat with Meta AI*, announced on 13 May 2026, takes that path: the Trusted Execution Environment in which the chat is processed is stateless by design, with no conversational history retained from one session to the next. The provider, in this architecture, has nothing to remember and therefore nothing to disclose under a Section 702 directive or its analogues. The cost in user-experience terms (the loss of memory-augmented continuity across sessions) and the longer-term question of whether confidential-memory architectures could relax this constraint without recreating the surveillance surface are addressed in Christakis and Swire (cited n. 11 above).

The convergence of memory, agentic capabilities, and health data integration documented in my companion study on the Health AI Agent Rush¹¹³ amplifies this concern. A user who has connected medical records, wearables, wellness apps, and years of conversational history to a single AI platform has created a comprehensive personal dossier that no intelligence agency could previously have assembled without extensive, multi-source surveillance. That dossier now sits, voluntarily constructed, on the provider's servers.

From passive memory to active agency: the implications for surveillance law

The surveillance implications extend further still when chatbots move from passive memory to active agency. Several providers are developing or have already deployed features that allow the AI to act autonomously on the user's behalf: browsing the web, writing and sending messages, contributing to shared documents or code repositories, managing calendars, executing purchases, and interacting with third-party services. Anthropic's "computer use" capabilities, OpenAI's operator and plugin ecosystem, and Google's agentic extensions for Gemini all point in this direction. Some implementations allow the AI to act without requiring explicit confirmation for each step, operating with delegated authority under general instructions from the user.

If government agencies can compel disclosure of records generated by these agentic actions, they gain access to a category of evidence whose specific combination of features, the user's natural-language intent, the system's autonomous execution across multiple external services, and the resulting real-world effects, held in a single provider repository, has fewer clear analogues in established surveillance doctrine than the familiar categories of communications and transactional records. A compelled disclosure of agentic action logs could reveal that the AI booked a flight, sent a message to a contact, edited a shared document, purchased specific items, or contributed code to a repository, all actions that the user may have authorised in general terms but never performed personally.¹¹⁴

Surveillance law has long distinguished communications, protected by warrant requirements, from transactional records available under lower thresholds, and separating the two within a single mixed record is routine work: emails, cloud files and app logs all bundle content with metadata, and providers and courts pull them apart as a matter of course. Agentic action logs do not unsettle that mature distinction so much as pose a narrower question at its edge. When an AI, acting on a general delegation rather than a specific instruction, sends a message, edits a document or makes a purchase, the record it leaves ties together three things the user did not separately produce: a natural-language intent, the model's autonomous execution across outside services, and a real-world effect. How the familiar categories attach to that bundle is the unsettled part: whether the fact that the agent sent a message is the user's communication or a transactional record, whether a machine-generated reasoning trace counts as content, and how delegated rather than direct authorship bears on either. The skill of classifying content against metadata is old; what is newer is the kind of record being classified.

2.7. Defence contracts and consumer data: does the dual role expose consumer chats?

Several major AI providers now operate consumer chatbot platforms used by hundreds of millions of people and, at the same time, supply models and infrastructure to United States defence and intelligence agencies. The natural question for a user of one of those platforms is whether that dual role exposes the consumer side to the government side: whether the existence of a Pentagon or intelligence contract opens a quiet route by which chatbot logs could reach those agencies, outside the protections that would normally apply to consumer data. The answer the public record supports is that no such route exists. Supplying a model for use on a classified government network is **a separate activity**, technically and legally, from running a consumer chatbot, and it follows the same separation that has long governed the cloud and software vendors who serve both markets. There is no public evidence that any provider has

¹¹³ Op. cit.

¹¹⁴ For the emergence of AI Agents and legal issues they raise see C. Busch, "[Consumer Law for AI Agents](#)", *German Law Journal*, 27(3), (2026), 112–125.

shared consumer chatbot logs with a defence or intelligence client, or that a defence contract has created any pathway from consumer data to government.

The dual role is now a prominent feature of the AI landscape, and its most visible recent illustration, the Anthropic-Pentagon dispute of 2025 and 2026, points the opposite way from a consumer-data pathway. The dispute was about the conditions under which Claude as a model could be deployed on government systems, not about consumer chatbot data. After Anthropic signed a large Department of Defense contract in 2025, it refused a later demand to drop its usage restrictions, holding to two red lines, no mass surveillance of people in the United States and no fully autonomous weapons. Anthropic CEO Dario Amodei stated publicly that the company “cannot in good conscience” accede to the Pentagon’s demand.¹¹⁵ The refusal escalated sharply: a federal directive to stop using the company’s technology, a “supply chain risk” designation, cancellation of the contract, and litigation that Anthropic won at first instance on First Amendment retaliation grounds¹¹⁶ and lost on a parallel motion on appeal.¹¹⁷ A provider publicly resisting the government over the terms on which its model may be deployed is the reverse of one quietly routing its users’ conversations to that government. When OpenAI then struck a Pentagon deal to fill the gap,¹¹⁸ it took on the same dual position the established cloud vendors have held for years.

The reason the answer is no lies in how this dual role has always been organised. In the cloud and government-services context the ordinary model is separation: distinct contractual instruments, segregated environments, separate access controls and, where they apply, separate accreditation regimes. Microsoft has for decades run commercial services such as Outlook.com and Bing alongside federal services delivered through Azure Government and Microsoft 365 Government; AWS GovCloud and Google Cloud’s federal offerings follow the same pattern. In none of these long-standing relationships is there any public allegation that the government-facing side of the business has given the government a route to commercial-customer data. AI providers entering the same dual role inherit that established pattern, and nothing on the public record shows their consumer systems wired to their government deployments.

What the AI setting changes is not that answer but the stakes of the underlying data. The records on the consumer side are not server logs or search queries but the intimate self-disclosure this study documents throughout: reasoning, plans, fears, health, relationships. The major providers already document the security of their government-facing environments in detail: Claude for Government and Claude on AWS GovCloud are authorised at FedRAMP High and Impact Level 5, and OpenAI publishes its FedRAMP authorisation data through a trust portal, as the established cloud vendors do. But that material is addressed to government customers and compliance reviewers; it certifies the protection of the government environment, not the consumer side. What is rarely, if ever, stated plainly for the ordinary consumer is the converse: that conversations held on the consumer platform are not reachable through the company’s government business. The recommendation is the modest one of making that specific assurance explicit, in writing and in language an ordinary user can follow. It does not require an outside auditor or system schematics. Because the value of these products rises and falls with the trust users place in them, stating the separation plainly serves the provider as much as the user.

2.8. Transparency reports: valuable but incomplete

Transparency reports are one of the few public windows onto how often, and under what authority, providers hand over data. They are not the tool an individual turns to in order to gauge personal risk:

¹¹⁵ See ASIS Online, [“Anthropic Refuses Pentagon Demand to Remove AI Security and Safety Guardrails”](#), 27 February 2026. ; CNN, [“Anthropic rejects latest Pentagon offer: ‘We cannot in good conscience accede to their request’”](#), 26 February 2026.

¹¹⁶ On 26 March 2026, Judge Rita Lin of the Northern District of California granted Anthropic’s motion for a preliminary injunction, finding that the government’s actions were “not designed to protect national security, but rather to punish Anthropic” and constituted “classic illegal First Amendment retaliation.” See CNBC, [“Anthropic wins preliminary injunction in DOD fight as judge cites ‘First Amendment retaliation’”](#), 26 March 2026.

¹¹⁷ On 8 April 2026, the U.S. Court of Appeals for the D.C. Circuit denied Anthropic’s request to temporarily block the Pentagon’s supply chain risk designation, finding that “the equitable balance here cuts in favor of the government” and citing “judicial management of how, and through whom, the Department of War secures vital AI technology during an active military conflict”. See CNBC, [“Anthropic loses appeals court bid to temporarily block Pentagon blacklisting”](#), 8 April 2026.

¹¹⁸ EFF, [“Weasel Words: OpenAI’s Pentagon Deal Won’t Stop AI-Powered Surveillance”](#), 6 March 2026.

for that a user is better served by the provider's privacy policy, retention settings and terms, by the user-facing disclosures recommended in Part 1, and by the press and court record. What transparency reports do, and what nothing else does at scale, is let regulators, courts, journalists and researchers see the systemic shape of compelled access. It is for that accountability function, not for individual risk assessment, that their granularity matters.

The AI-specific reporters. OpenAI and Anthropic deserve credit for publishing dedicated AI-specific government-request reports. OpenAI's first report, for January–June 2023, listed 6 non-content requests, no content requests, and no emergency requests.¹¹⁹ By July–December 2025, OpenAI reported 224 non-content requests, 75 content requests, and 10 emergency requests, with data disclosed in 146, 62, and 1 respectively. Anthropic's figures remain much lower: one non-content request and no content or emergency requests in July–December 2024, and zero across all three categories in January–June 2025.¹²⁰ However, these low figures should not be over-read. Anthropic's user base is still growing, and the reporting history of both providers is too short to establish any settled level, still less a stable one, against which later numbers could be judged.

The aggregated reporters. Google and xAI fold AI-related disclosure into their general corporate transparency reporting. Google reports 287,014 global requests for user information for January–June 2025 across all its services,¹²¹ and X reports 20,925 for the second half of 2024 across the X and Periscope ecosystem. Neither figure can be broken down to show how many demands touched a chatbot. That is partly a reporting choice, but it is also a feature of how legal process works against a multi-service provider: a demand is usually directed at a user's account, which spans every service that account subscribes to, not at a single product, so for many demands there is no one service to assign them to. A provider could perhaps count content productions by service, where content from a given service is in fact produced, but a clean per-service breakdown of demands as a whole would be a heavy and often not meaningful exercise. The limitation for a user or researcher is nonetheless real: from the published aggregate, there is no way to tell how many demands reached the intimate interactions of chatbot users. It is better understood as an artefact of account-level legal process than as a deliberate burial of separable figures.

The transparency void. DeepSeek is the outlier. As of April 2026, no dedicated official transparency report quantifying government requests for consumer chatbot logs has been identified. For a service that states it directly stores personal data in the PRC, subject to Chinese national security and intelligence cooperation laws, that absence is especially important. This study does not undertake a substantive comparative analysis of Chinese national security and intelligence cooperation law. The point made here is narrower and concerns the publicly observable transparency posture only: a consumer service that processes hundreds of millions of conversations and that publishes no quantified report on government access leaves users without the most basic information needed to assess that risk, regardless of the underlying legal framework. The asymmetry should therefore not be read as suggesting that the United States is uniquely problematic; it reflects **the opposite**: U.S. companies disclose enough to allow scrutiny, while the absence of comparable reporting by DeepSeek prevents even a basic public assessment.

The growth trajectory.

The current volume of government requests for chatbot data is small compared to legacy technology platforms. But the trajectory is the same. Google received 13,424 requests in January–June 2010. Fifteen years later, that number had reached 287,014: an increase of over 2,000%. OpenAI received 6 government requests for user data in its first reporting period (H1 2023) and 309 in H2 2025, a more than fifty-fold increase over two and a half years, although part of this growth also tracks the substantial expansion of OpenAI's user base over the same period. The comparison is not predictive. Google's trajectory reflects the maturation of a legacy platform ecosystem over fifteen years. OpenAI's increase

¹¹⁹ OpenAI, Report on Government Requests for User Data (July–December 2025). Available at: <https://cdn.openai.com/trust-and-transparency/report-2025h2-government-requests-for-user-data.pdf>.

¹²⁰ Anthropic, Government Requests Report (January–June 2025) and Government Requests Report (July–December 2024). Available at: <https://assets.anthropic.com/m/604a7603983db0b9/original/Anthropic-Government-Requests-Report.pdf>.

¹²¹ Google Transparency Report, "Global requests for user information" (showing 13,424 requests for January–June 2010 and 287,014 for January–June 2025). Available at: <https://transparencyreport.google.com/user-data>.

starts from a tiny baseline and covers only two and a half years. But the direction seems to be that once a service becomes socially central, government requests follow. Whether the eventual volume of government requests for chatbot data reaches the levels seen with legacy electronic-communications platforms is genuinely uncertain. Some of the factors that drive investigators toward email and social media (communications between identified parties, networks of contacts, posts visible to others) apply less directly to chatbot conversations, while other factors that the present study has documented (the unusually intimate nature of disclosure, the presence of memory and agentic-action logs) do not have close analogues in earlier platforms. What the early data does show, with chatbot adoption now estimated at well over one billion weekly chatbot interactions/users across major services¹²², is a trajectory steep enough, and a set of access categories already varied enough (subpoena, search warrant, emergency disclosure, the first federal reverse prompt warrant), to make the question of how these requests are reported a present-tense concern rather than a future one.

Transparency reports are a valuable accountability instrument, but they are incomplete in ways that matter. Three improvements would help, and it is worth being clear which is specific to chatbots and which is not. The first is general rather than AI-specific: providers that run several service lines could break their corporate aggregates down by service, so that a demand concerning Gemini or Grok can be told apart from one concerning Maps, Search or an X account. The same argument has long been made about those other services and is no stronger for chatbots; it has simply not been acted on. It is also more work for a multi-service provider than for a single-product one: OpenAI's and Anthropic's reports are chatbot-specific only because a chatbot is essentially all they offer, so for them nothing needs breaking out, whereas a single demand to a multi-service provider may span several products and has to be attributed before it can be counted. The second and third improvements are the ones particular to this setting, because they concern demands that do not arise for a mailbox or a map. Providers could separately report reverse-prompt demands, directed at the content of users' prompts rather than at a named account, which the first such federal warrant has shown to be no longer hypothetical. And they could separately identify demands targeting memory or personalisation features, since a memory file is a structured longitudinal profile rather than a set of discrete messages and does not aggregate cleanly into a transcript count. None of the three requires disclosing categories the law prohibits. Together they would let regulators, courts, journalists and researchers see not just how often the door was opened but which rooms the government entered.

¹²² While no official census exists, the figure accounts for approximately 900 million weekly ChatGPT users, 185 million weekly Meta AI users, and several hundred million across Gemini, Claude, and others. See e.g. <https://resourcera.com/data/artificial-intelligence/chatgpt-users/>.

Table 3. Transparency reporting and disclosure metrics

Dimension	ChatGPT (OpenAI)	Claude (Anthropic)	Gemini (Google)	Grok (xAI)	DeepSeek
Dedicated AI-specific government-request report	Yes. Published since 2023.	Yes. Published since 2024.	No. Google-wide reporting only (since 2009).	No. X-wide reporting only.	No dedicated report identified.
AI-specific identifiability of requests	High. Requests concern OpenAI services specifically. Distinguishes content, non-content, emergency.	High. Reports concern Anthropic services specifically. Same three-category distinction.	Low. Gemini statistically buried in Google-wide reporting (287,014 total requests, H1 2025).	Low. Grok buried in X/Periscope aggregate (20,925 total requests, H2 2024).	None publicly quantified.
Latest public figures	H2 2025: 224 non-content, 75 content, 10 emergency requests. Data disclosed in 146, 62, and 1 respectively. National security: 0–249.	H1 2025: 0 content, 0 non-content, 0 emergency. H2 2024: 1 non-content request. National security (2024): 0–99.	H1 2025: 287,014 global user-information requests across all Google services.	H2 2024: 20,925 X/Periscope information requests.	No quantitative disclosure.
Reverse prompt disclosure	Not separately reported.	Not separately reported.	Google previously published geofence warrant supplement; no AI equivalent.	Not separately reported.	Not applicable (no reporting).

Note: Assessments based on publicly available materials as of April 2026.

2.9. Conclusions and tentative recommendations

Six main conclusions emerge from this analysis.

First, the legal architecture of compelled access is already operating against chatbot providers in ordinary criminal cases. Subpoenas, court orders, search warrants, and preservation requests under the established US framework, and the parallel instruments in other jurisdictions (while waiting for the implementation of E-Evidence in the EU in August 2026), are being served on AI providers and being complied with according to the same basic ladder that has long governed access to electronic communications. The frontier categories that have emerged in the past eighteen months, in particular the first publicly reported federal reverse prompt warrant served on an AI provider in October 2025, remain rare events at the time of writing. Whether they multiply, and at what rate, is a question the early evidence cannot decisively answer; what the early evidence does establish is that the framework is operationally available for use against chatbot data and that several of the legal questions raised by its application to this category of data remain unsettled.

Second, the surveillance significance of consumer chatbot logs is in some respects continuous with that of earlier forms of electronic communication and in some respects distinct. The observation that an email thread or a sequence of search queries can reveal a great deal about a user's reasoning, plans, and personal circumstances has been made many times before, and was made of those earlier categories at

their own moments of regulatory attention. What this study has argued to be more specifically distinctive in the chatbot context is the continuous-record form of conversational interaction, the structured longitudinal profile that memory features create where they are enabled, and the emerging category of agentic action logs that bind natural-language intent, autonomous execution and real-world effect in a single record. The established categories of communications and transactional records carry over, and providers and courts apply them to mixed records as a matter of routine. What is newer is not the classification work but the kind of record being classified, in particular the agentic action log and the synthesised memory profile. How the established categories apply to records of that kind is a question the law has not yet had occasion to settle.

Third, transparency practices vary across the industry, also due to the profile of the relevant companies. OpenAI and Anthropic, which are AI startups, publish dedicated AI-specific government-request reports, which is unsurprising given that AI is the main (if not the only) service they offer; Google and xAI report AI-related requests within broader corporate transparency reporting that does not separately identify chatbot-specific demands. These differences are not merely presentational. They determine whether regulators, courts, journalists and researchers can see how often, and under what authority, chatbot records are sought, preserved or produced. As Section 2.8 sets out, the case for more granular reporting rests on that accountability function rather than on any claim that individual users consult these figures to gauge personal risk; and only part of it is specific to chatbots, since breaking corporate aggregates down by service is a general point while separate reporting of reverse-prompt and memory-targeted demands is particular to this setting.

Fourth, in the law-enforcement context the open question is whether ordinary criminal process could be used to run a bulk reverse search of a chatbot database, and it turns on a single requirement: particularity. A narrow warrant uses a specific communication the investigators already hold as a fingerprint to identify the person behind it: it names what is sought, and the search only resolves who sent it. A bulk warrant does the reverse, describing a category of content, a topic or theme, and asking the provider to surface everyone whose conversations fall within it, so that the search itself defines both what and who. The first is the particular description a warrant is meant to contain; the second is the open-ended sweep the keyword and geofence cases treat as a possible general warrant. The first known reverse-prompt warrant, served on OpenAI in October 2025, was of the first kind: investigators already held two specific prompts the suspect had shared during undercover contact, and the warrant used them to reach a single account. A bulk warrant would begin instead from a content pattern, with no specific communication and no prior suspect, and return an open-ended class of users. As Section 2.4 explains, geofence, reverse keyword and reverse-prompt warrants are variants of a single technique, compelling a provider to search its whole database to find the unknown people who match a pattern the police specify; the bulk reverse-prompt warrant is that technique applied to prompts, and it raises the same particularity question. The constitutional rules the courts settle for one will largely govern the others, so the principles developed in those cases apply to reverse-prompt warrants directly, not merely by analogy.

Fifth, the same kind of search could in principle be sought by the other route, through intelligence law rather than criminal process, and here the most aggressive hypothesis explored in this chapter, that intelligence agencies might compel a provider to search its whole consumer database for any user whose conversations match a content pattern, regardless of whether that user has been identified as a target, faces real statutory obstacles under US surveillance law as it currently stands. Section 702 authorises the targeting of non-US persons reasonably believed to be abroad, using tasked selectors associated with those targets, rather than generalised content searches across a provider's user base, and the post-RISAA prohibition on resuming 'abouts' collection strengthens the argument that analogous content-based searches would face serious obstacles. To that extent the analysis is reassuring about the US framework as it operates today. What it does not exclude is a narrower and more realistic hypothesis: a targeted demand, against a validly targeted non-US person abroad, for the record types this chapter has described, memory profiles and agentic logs. The open question is therefore how exactly the necessity and proportionality tests of Executive Order 14086, framed around signals intelligence collection, apply to a structured profile that synthesises years of a person's life, a question for the bodies that administer and periodically review the Data Privacy Framework to work out in practice.

Sixth, for DeepSeek both of these questions, law-enforcement access and intelligence access alike, fall under a body of law this study has not examined: the Chinese framework of law-enforcement, national-security and intelligence-cooperation obligations that would govern the consumer chatbot data DeepSeek holds. That analysis lies beyond the scope of the present work and needs dedicated treatment in its own right. The gap is important because DeepSeek combines the regime examined least here with the least transparency of any provider considered. Where OpenAI and Anthropic publish dedicated reports, and the multi-service US providers report in the aggregate, DeepSeek publishes no quantified report on government access to its consumer chatbot at all. The provider operating under the framework about which the public can infer the least therefore also offers the least to go on, and that is a combination that should attract scrutiny.

Tentative Recommendations on Compelled Government Access

It is worth being clear at the outset about what these recommendations can and cannot achieve. When it comes to government access to data, providers are not the principal actors. The terms on which a provider must disclose user data are set by statute, and by the courts that interpret it; a provider served with valid legal process must, in the ordinary case, comply. The questions that matter most, what the standard for compelled disclosure should be, which categories of data deserve which protection, and whether reverse searches of prompt databases are constitutional, are questions for legislators, who write the rules, and for judges, who apply them, not for the companies that receive the demands. The providers' room for manoeuvre is correspondingly narrow. The recommendations that follow are therefore modest by design: each is confined to what a provider can do on its own initiative, within the law as it stands, and without taking positions on the policy questions that lie beyond the scope of a study of this kind. They will not change the legal standard for compelled access, which only legislators and courts can do; what they can do is make the existing position clearer to users and easier to scrutinise, and keep providers' own practices abreast of the new categories of demand this chapter has identified. With those limits in mind, we suggest three things providers might usefully do from their side.

Tentative Recommendations on Compelled Government Access

#	Recommendation
6	Improve transparency reporting on compelled access to chatbot data. Three improvements would help regulators, courts, journalists and researchers see the shape of compelled access to chatbot data. The first is general rather than AI-specific: providers that run several service lines could, to the extent possible, break their corporate aggregates down by service, so that a demand concerning Gemini or Grok can be distinguished from one concerning Maps, Search or an X account. This is more work for a multi-service provider than for a single-product one, since a single demand may span several products and must be attributed before it can be counted; OpenAI's and Anthropic's reports are chatbot-specific only because a chatbot is essentially all they offer, so for them nothing needs breaking out. The second and third improvements are particular to this setting, because they concern demands that do not arise for a mailbox or a map. Where legally permitted, providers could separately report reverse-prompt demands, which are directed at the content of users' prompts rather than at a named account and which the first such federal warrant has shown to be no longer hypothetical. And they could separately identify demands targeting memory or personalisation features, since a memory file is a structured longitudinal profile rather than a set of discrete messages and does not aggregate cleanly into a transcript count. None of the three requires disclosing categories the law currently prohibits.
7	Confirm publicly that the commitment to resist overbroad demands reaches bulk reverse searches. Several providers already state that they review demands carefully and resist those that are overbroad or defective, and for most of the new categories this chapter identifies that general commitment already does the work. The exception worth a specific public statement is the demand this chapter treats as the most dangerous and the most novel: a bulk or content-

	<p>pattern search across the prompt database, with no nexus to an identified user. A provider stating plainly that it will resist demands of that shape, while making clear the statement is illustrative and does not narrow its general commitment, gives courts and users a concrete expectation at the point where the established posture is least settled.</p>
8	<p>Treat memory and agentic action logs as distinct legal-process categories. Providers already separate content from transactional data as a matter of routine; the recommendation here is narrower, that they name memory profiles and agentic action logs as distinct record types in legal-process review, transparency reporting, retention design and user-facing explanations, and disclose how each is handled. Memory files are structured longitudinal profiles; agentic logs bind natural-language intent, autonomous execution and real-world effect in a single record. Treating them explicitly, rather than folding them into the ordinary transcript category, makes visible a distinction the established framework otherwise handles only implicitly.</p>

3. Discovery: When Chatbot Conversations Become Evidence

In December 2025, a federal magistrate judge ordered OpenAI to produce 20 million de-identified ChatGPT conversations to opposing counsel in a copyright case.¹²³ Three months later, according to MLex reporting, the production fight expanded to a newly identified repository of 88 million anonymous ChatGPT conversations.¹²⁴ The conversations belonged to people who had no connection to the litigation.

The discovery phase of the New York Times litigation¹²⁵ has turned what began as a copyright dispute into something broader: a test of whether consumer chatbot conversations stay confidential once litigation reaches for them. That case is the dramatic version of the question, because the conversations sought belong to millions of people who are strangers to the suit and are demanded from the provider in bulk. It is not, however, the version most likely to reach an ordinary person. The same treatment that makes a stranger's conversation producible at scale makes a litigant's own conversations producible in their own case, so that the record a person builds up in a chatbot, about a divorce, an employment dispute, a custody fight or a business quarrel, can be called for in discovery and read by the other side. This chapter is about both faces of one development: courts have begun to treat chatbot conversations as ordinary evidence, whether they belong to strangers and are swept up by the million, or to the litigant and are sought for what they reveal.

As argued in Part 1 of this study, chatbot conversations are unusually sensitive. They often contain emotional disclosures, legal and medical discussions, business plans, code, drafts, payment-related information, credentials, and highly individualised factual narratives. OpenAI itself stated in November 2025 that users entrust ChatGPT with “some of the most personal parts of their lives”, and that hundreds of millions of people use the service each week.¹²⁶ Once litigation reaches conversations of this kind, whether in bulk or one account at a time, the legal issue becomes whether ordinary AI conversations remain meaningfully confidential under that pressure.

This chapter argues that discovery of chatbot conversations erodes confidentiality through four analytically distinct stages: preservation, production, review, and downstream risk. Preservation already weakens confidentiality because data that would otherwise be deleted must remain available. Production broadens the circle of access. Review by outside counsel, experts, and litigation vendors creates a qualitatively new exposure. And downstream risk persists even under protective orders because modern litigation depends on many actors, systems, and repositories that can fail.

The chapter proceeds in eight steps. First, it reconstructs what happened procedurally in the NYT case. Second, it explains OpenAI's privacy defences and why they did not prevail. Third, it separates the two questions that the phrase “AI privilege” runs together, whether any privilege shields a provider's records of consumer conversations, and whether using an AI tool costs a litigant the protection their own preparation would otherwise carry and shows that current law leaves the user with no dependable protection on either: no privilege of the first kind exists, in the US or in other major systems, while the second question remains unsettled, split between decisions such as Heppner and Warner. Fourth, it explains why de-identification could not fully solve the problem. Fifth, it examines why protective orders reduce risk but do not eliminate it, drawing on precedents involving discovery misuse, public

¹²³ Reuters, “[OpenAI loses fight to keep ChatGPT logs secret in copyright case](#)” (3 December 2025).

¹²⁴ Emma Whitford and Samuel Rubinfeld, “OpenAI must produce millions more conversations in US copyright cases,” MLex, 10 March 2026.

¹²⁵ The New York Times Co. v. Microsoft Corp. and OpenAI, Inc., complaint filed 27 December 2023, S.D.N.Y. See also Reuters, “[Judge explains order for New York Times in OpenAI copyright case](#)” (4 April 2025).

¹²⁶ OpenAI, “[Fighting the New York Times' invasion of user privacy](#)” (12 November 2025). See also OpenAI, “[How we're responding to The New York Times' data demands in order to protect user privacy](#)” (5 June 2025, updated 22 October 2025).

docket exposure, and cyberattacks affecting law firms and e-discovery vendors. Sixth, it asks whether a New York Times-style order could arise in Europe, through the disclosure regime of England and Wales, the IP Enforcement Directive, national procedural mechanisms or regulatory investigations. Seventh, it sets out the several other routes by which chatbot conversations now become evidence, from bulk production of strangers' logs to a litigant's own conversations in an ordinary case. Eighth, it explains why such discovery efforts are likely to persist and multiply. The chapter ends with a series of conclusions and recommendations.

3.1. What exactly happened in the New York Times discovery battle?

The underlying case

On 27 December 2023, *The New York Times* sued OpenAI and Microsoft in the Southern District of New York, alleging, among other things, copyright infringement arising from the use of Times content in the development and operation of OpenAI's models. In spring 2025, Judge Sidney H. Stein allowed key copyright claims to proceed, setting the stage for more aggressive discovery concerning both model behaviour and real-world user interactions.¹²⁷

The plaintiffs' theory was not limited to abstract training-stage copying. They also argued that OpenAI's products could reproduce or closely paraphrase protected works in response to user prompts. That made logs of user prompts and model outputs potentially attractive evidence. If the plaintiffs could show that actual users were eliciting infringing outputs at scale, the logs could serve at least three litigation functions: proving regurgitation, testing market-substitution theories, and rebutting OpenAI's claim that some of the plaintiffs' examples had been manufactured or "hacked".

From preservation to production

The dispute began with a preservation fight. In June 2025, OpenAI appealed a court-ordered data preservation requirement entered at the plaintiffs' request. OpenAI argued that the order required it to keep user chats, including chats that users had deleted, in a manner inconsistent with its privacy commitments. OpenAI's public statement of 5 June 2025 described the order as forcing the company to retain consumer ChatGPT and API content going forward, including content that would otherwise be deleted within 30 days.¹²⁸

At that stage, the confidentiality problem was already real. A preservation order does not yet disclose material to opposing counsel, but it interrupts ordinary deletion and increases the amount of sensitive material held under legal hold. OpenAI publicly stressed that the preservation order affected ChatGPT Free, Plus, Pro, Team, and API users without a Zero Data Retention agreement, while not affecting ChatGPT Enterprise, ChatGPT Edu, or API customers using Zero Data Retention endpoints. This point is important: discovery exposure is shaped not only by doctrine but by product architecture. Services built around stronger segregation, enterprise controls, or genuine zero-retention commitments are materially less vulnerable to bulk litigation capture.

The geographic scope of the preservation obligation also evolved. In its 22 October 2025 update, OpenAI stated that it was no longer under an order to retain consumer ChatGPT and API content indefinitely, that its obligations under the earlier order had ended on 26 September 2025, and that it was no longer required to retain new user data going forward or "any conversations originating from the European Economic Area, Switzerland, or the United Kingdom". OpenAI added, however, that the Times continued to demand preservation of a specific set of historical user data from April to September 2025, which OpenAI said would remain securely stored and accessible only to a small, audited legal and security team. This clarification matters.¹²⁹ The preservation dispute did not simply impose a permanent, undifferentiated legal hold on all users everywhere. Its scope changed over time, and

¹²⁷ Reuters, "[Judge explains order for New York Times in OpenAI copyright case](#)" (4 April 2025).

¹²⁸ Reuters, "[OpenAI appeals data preservation order in NYT copyright case demanding no one delete any user chats](#)" (6 June 2025).

¹²⁹ OpenAI, "[How we're responding to The New York Times' data demands in order to protect user privacy](#)", update of 22 October 2025.

OpenAI publicly distinguished European-originating conversations from the continuing US litigation hold. But the episode remains significant because it demonstrates that litigation can temporarily override ordinary deletion expectations, create legally segregated corpora of user conversations, and make the practical protection of privacy depend on product tier, retention architecture, geography, and the evolving terms of court orders.

The conflict then escalated **from preservation to production**. In November 2025, Reuters reported that OpenAI had asked a federal judge to reverse an order requiring it to turn over 20 million anonymised ChatGPT conversations to *The New York Times* and other news plaintiffs.¹³⁰ OpenAI described the order as a grave intrusion into the privacy of users unrelated to the litigation. Its public statement of 12 November 2025 said the Times was demanding 20 million private ChatGPT conversations, after first trying to stop users from deleting chats and after earlier seeking up to 1.4 billion private conversations.¹³¹

The European dimension must be emphasized here. In its FAQ accompanying the November 2025 public statement on the 20-million-conversation production order, OpenAI expressly addressed the question whether the court order violated the GDPR or users' rights under European or other privacy laws. Its answer did not state that European users were outside the scope of the order. Instead, OpenAI stated that it was "taking steps to comply at this time because we must follow the law," while adding that "The New York Times' demand does not align with our privacy standards" and that this was why OpenAI was challenging it.¹³² The formulation is significant: unlike OpenAI's October 2025 update on the earlier preservation order, which stated that the company was no longer required to retain conversations originating from the EEA, Switzerland, or the United Kingdom, the November 2025 production-order FAQ treated GDPR and European privacy rights as live concerns rather than as categorically excluded from the dispute.

On 3 December 2025, Magistrate Judge Ona T. Wang ordered OpenAI to produce the 20 million de-identified logs, rejecting the company's privacy objections and accepting the plaintiffs' relevance arguments.¹³³ OpenAI later confirmed publicly that the district judge had ruled on appeal and that the data had undergone a de-identification process intended to remove or mask personally identifying and other private information before disclosure under strict access controls.¹³⁴

On 5 January 2026, Judge Sidney H. Stein affirmed Judge Wang's orders.¹³⁵ The district court held that users' privacy interests had been adequately accounted for and that the magistrate judge had not clearly erred in requiring production of the entire 20 million de-identified sample rather than OpenAI's narrower search-term proposal. The court also emphasised that the conversations were information that users had voluntarily disclosed to OpenAI and that no party argued OpenAI possessed them unlawfully.¹³⁶ This point is central to understanding why OpenAI's privacy defence failed. The court did not treat the logs as categorically immune from discovery merely because they were private or sensitive. Rather, it treated user privacy as a proportionality factor to be balanced against relevance, de-identification, access controls, and the protective order. The court's emphasis on users' voluntary disclosure to OpenAI also illustrates a broader problem: under existing U.S. doctrine, information shared with a service provider is often less protected than users intuitively expect. But the ruling should be understood first as a discovery proportionality decision, not as a definitive constitutional holding on the Fourth Amendment status of chatbot logs.

¹³⁰ Reuters, "[OpenAI fights order to turn over millions of ChatGPT conversations](#)" (12 November 2025).

¹³¹ OpenAI, "[Fighting the New York Times' invasion of user privacy](#)", 12 November 2025.

¹³² OpenAI, "[Fighting the New York Times' invasion of user privacy](#)", 12 November 2025, FAQ. "The same FAQ described the order as concerning 20 million randomly sampled consumer ChatGPT conversations from December 2022 to November 2024 and excluded ChatGPT Enterprise, ChatGPT Edu, ChatGPT Business, and API customers from the impacted products, but did not state that EEA, Swiss, or UK consumer users were categorically excluded from the production sample."

¹³³ *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:24-md-03109 (S.D.N.Y.), Order of Magistrate Judge Ona T. Wang denying OpenAI's motion for reconsideration (3 December 2025). I was not able to find a direct link to the full text of Judge Wang's order itself (court orders in SDNY are typically available only through PACER or ECF). However, I found very detailed reporting on this judgment that I use in my analysis below.

¹³⁴ OpenAI, "[The facts concerning the New York Times's lawsuit against OpenAI](#)" (update dated 16 December 2025).

¹³⁵ *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:24-md-03109 (S.D.N.Y.), Order of District Judge Sidney H. Stein (5 January 2026). Here also I was not able to find a direct link to the full text of Judge Stein's order itself as it is again on PACER/ECF behind a paywall. But I use analytical reporting by several sources including the detailed presentation by Roma Patel mentioned in the next footnote.

¹³⁶ Roma Patel, "[When Chats Become Evidence: Court Affirms Order Requiring OpenAI to Produce 20 Million De-Identified ChatGPT Logs](#)" (15 January 2026), Data Privacy + Security Insider.

The March 2026 expansion to 88 million conversations

The discovery battle did not end with the 20 million-log sample. In February and March 2026, the dispute widened sharply after plaintiffs learned, through the late-January deposition of OpenAI engineer Vinnie Monaco, of a much larger repository of searchable, “anonymised” ChatGPT conversations that had not previously featured in the discovery record. According to contemporaneous MLex reporting, plaintiffs told the court that OpenAI had maintained at least 78 million searchable anonymised conversations sourced in the spring of 2023 or earlier, and later identified an additional 10 million, bringing the total to 88 million conversations. Plaintiffs argued that the existence of this repository contradicted their earlier understanding that OpenAI lacked the capacity to run comprehensive searches across relevant chatbot outputs. They further contended that the newly identified logs were especially probative because they appeared to pre-date OpenAI’s later filter designed to suppress “regurgitation” of copyrighted works. The logs allegedly surfaced in connection with “Project Giraffe,” an internal OpenAI effort to test or improve output filtering.¹³⁷

At a 11 February 2026 conference, Magistrate Judge Ona T. Wang ordered OpenAI to clarify urgently the scope and status of this repository, including whether the logs still existed, whether OpenAI had searched plaintiffs’ works across them, and whether other similar output repositories existed.¹³⁸ OpenAI resisted the plaintiffs’ characterisation, arguing that the material did not demonstrate a generalised ability to search all output logs, but rather concerned a more limited “slice” of data collected to test a filter.

The matter culminated on 10 March 2026, when MLex reported that Judge Wang ordered OpenAI to turn over the 88 million recently discovered anonymous ChatGPT conversations, in addition to the earlier 20 million-log sample, while the parties were to discuss the particulars of the handover, including user-privacy guardrails, in a confidential settlement conference.¹³⁹ OpenAI publicly denounced the ruling as a “fishing expedition,” arguing that plaintiffs had already combed through 20 million conversations without finding support for their claims and were now demanding access to tens of millions more.

Because the underlying orders and conference materials are not fully public, the precise relationship between the earlier 20-million sample and the later 88-million repository should be stated cautiously. The key point is not whether the later set should be treated as a simple numerical enlargement of the first, but that the discovery perimeter expanded dramatically once additional searchable conversation repositories came into view.

Whether one emphasises plaintiffs’ theory of relevance or OpenAI’s privacy objections, the March 2026 development is crucial because it shows that once a court accepts the discoverability in principle of large-scale chatbot logs, the scope of production may expand dramatically as new repositories, older datasets, or previously undisclosed search capabilities come to light. The central lesson is structural: chatbot-log discovery is not a fixed event, but a process that can widen materially as litigation progresses.

3.2. OpenAI’s privacy defences and why they failed

3.2.1. The substance of OpenAI’s privacy defence

OpenAI’s position had several strands.

First, it argued that the plaintiffs’ requests were drastically overbroad relative to the needs of the case. In its 10 December 2025 memorandum, OpenAI described the compelled disclosure of 20 million private conversations as one of the largest court-ordered transfers of private personal information in U.S. litigation and argued that the overwhelming majority of those logs lacked any particularised relevance to the plaintiffs’ claims. OpenAI proposed a less intrusive alternative: use search terms, n-

¹³⁷ Emma Whitford, “OpenAI faces possible expanded ChatGPT log production as US discovery closes,” MLex, 11 February 2026.

¹³⁸ Ibid.

¹³⁹ Emma Whitford and Samuel Rubinfeld, “OpenAI must produce millions more conversations in US copyright cases,” MLex, 10 March 2026.

gram matching, and metadata-based methods to identify only those records specifically linked to plaintiffs' works while relying on non-content signals for broader statistical analysis.¹⁴⁰

Second, OpenAI argued that de-identification was being overstated. Its filings stressed that removing names, contact information, and numeric identifiers does not eliminate the privacy significance of conversations that still contain intimate, proprietary, or otherwise sensitive substance. The company's December 2025 memorandum used concrete examples and stated that removing identifiers from logs discussing "details of their finances, immigration status, trading strategies, or investigative journalism" does not make those records "no longer private, sensitive, confidential, or privileged".

Third, OpenAI argued that compelled disclosure would expose unrelated users to outside actors. In its November 2025 public statement, the company warned that if the Times succeeded, user data could be made available to "third parties, including the Times' lawyers and paid consultants."¹⁴¹ That argument highlighted a reality of modern litigation: once data enters discovery, it is often handled not only by opposing counsel but also by expert witnesses, contract reviewers, vendors, and secure-review administrators.

Fourth, OpenAI relied on user expectations shaped by its deletion and retention promises. In June 2025, the company emphasised that deleted ChatGPT conversations and API content are ordinarily removed within 30 days and portrayed the preservation order as undermining long-standing privacy commitments. The point was that consumer-facing deletion controls lose much of their practical meaning if litigation can compel the provider to retain deleted material at scale.

Finally, OpenAI made a broader systemic argument: discovery demands of this kind could chill trust in consumer AI and create asymmetries between providers with stronger enterprise segregation and providers whose consumer products are deeply integrated into everyday life.

3.2.2. Why the defence did not prevail

The reason OpenAI lost is not that its privacy concerns were trivial. Rather, the courts treated them as insufficient under ordinary U.S. discovery doctrine. Federal Rule of Civil Procedure 26 permits discovery regarding any nonprivileged matter that is relevant to a party's claim or defence and proportional to the needs of the case.¹⁴² Once the courts accepted the plaintiffs' arguments that the logs could show both infringement and non-infringing uses relevant to fair-use defences, the dispute shifted to proportionality rather than categorical immunity.

Where sensitive material is relevant and proportional, U.S. civil procedure ordinarily manages confidentiality through protective orders under Rule 26(c), including access restrictions, attorneys'-eyes-only designations, secure review protocols, and limits on onward use. That procedural architecture explains the outcome: the privacy issue was translated into conditions of production, not a categorical refusal of discovery.

Judge Stein's January 2026 affirmance reflects that logic. The court concluded that Judge Wang had adequately balanced users' privacy interests against the discovery value of the logs and that she was not required to adopt OpenAI's narrower search-term approach merely because it would have been less intrusive. The court also relied on the fact that users had voluntarily shared the information with OpenAI, which, under established third-party doctrine, significantly reduces the expectation of privacy. In other words, the court treated privacy as a factor to be weighed, not as a barrier that would stop the production altogether.

There was also a second, deeper doctrinal problem for OpenAI: ordinary chatbot conversations are private in a colloquial sense, but they are not generally privileged in the legal sense.

¹⁴⁰ OpenAI, [Memorandum of Law](#), filed 10 December 2025, ECF No. 935, In re OpenAI, Inc. Copyright Infringement Litigation, at 4, 11–12.

¹⁴¹ OpenAI, ["Fighting the New York Times' invasion of user privacy"](#) (12 November 2025).

¹⁴² [Fed. R. Civ. P. 26\(b\)\(1\)](#).

3.3. “AI privilege”

The discovery fight has put the phrase “AI privilege” into wide circulation, but the phrase is carrying two different questions that are better kept apart.

The first is whether any evidentiary privilege shields a provider’s records of a user’s conversations from discovery by someone else, the question the *New York Times* dispute raises, where the conversations sought belong to people who are strangers to the litigation.

The second is whether a litigant who uses a consumer AI tool to prepare their own case keeps the protection such preparation can carry, chiefly work-product, or instead leaves behind a discoverable record of their litigation thinking, the question raised by the Heppner and Warner decisions.

What separates them is whose protection is in play. A privilege belongs to the person who confided, not to the company that holds the record of the confidence. The first question therefore asks *whether a new privilege* should be created to protect records the provider holds; the second asks *whether the protection a litigant’s own preparation can already carry survives the involvement of an AI tool*. The two share a root, the law has not settled how AI conversations map onto the categories of confidentiality, and they share a remedy, as the conclusion of this section explains. But they are different doctrines, asserted by different people, in different settings, and treating them as one blurs both.

3.3.1. *The first question: no privilege shields a provider’s records of consumer AI conversations*

Federal Rule of Evidence 501 does not create a freestanding privilege for communications with AI systems. It instead provides that privilege in federal proceedings is governed by the common law “as interpreted by United States courts in the light of reason and experience,” except where state law supplies the rule of decision.¹⁴³ U.S. law therefore recognises specific privileges, such as attorney-client privilege, doctor-patient protections under particular legal regimes, clergy privilege in some settings, and related doctrines developed by statute or common law. But there is no general privilege for ordinary consumer AI conversations.

This gap matters because many users experience a chatbot as a confidant, brainstorming partner, quasi-therapeutic listener, substitute for a search engine, or even a rough stand-in for a lawyer or doctor. Yet, unless a recognised privilege attaches through some separate doctrine, the law may still treat the resulting exchanges as discoverable records held by a third-party provider. The gap between felt confidentiality and legal privilege is one of the deepest vulnerabilities exposed by the OpenAI discovery dispute.

Two clarifications keep this precise. First, the absence of a privilege is the ordinary condition of most human exchange, not a special deprivation: conversations with friends, colleagues and most other confidants enjoy none either. What is striking is not that chatbot conversations lack a privilege but that users increasingly confide in them in the register of the few relationships that carry one, the lawyer, the doctor, the therapist, the member of the clergy, the spouse. Second, the lack of a privilege does not mean the conversations are unprotected. Provider-held conversations carry statutory protection under the Stored Communications Act, and their status under the Fourth Amendment against compelled government access is a live and contested question, as Chapter 2 sets out. But those are protections against the state, not evidentiary privileges, and they do nothing to shield a provider’s records from discovery in litigation between private parties. The gap the discovery dispute exposes is the absence of a privilege of the kind that blocks disclosure regardless of relevance.

3.3.2. *Sam Altman’s proposal, and the trouble with a privilege the provider would hold*

The idea of an “AI privilege” entered mainstream debate when Sam Altman publicly warned in July 2025 that conversations with ChatGPT do not enjoy the same legal confidentiality as communications with therapists, doctors, or lawyers. In discussing the issue, he suggested that some form of AI-specific

¹⁴³ [Fed. R. Evid. 501](#).

privilege or equivalent legal protection may eventually be needed for especially sensitive interactions.¹⁴⁴ The timing was significant: Altman's intervention came just weeks after the May 2025 preservation order in the NYT case, meaning that the CEO of OpenAI was publicly calling for an AI privilege while his own company was fighting, and losing, a discovery battle that would have been resolved by exactly such a privilege.

That intervention was important because it came from the chief executive of the company at the centre of the largest chatbot-discovery fight to date. But, as a legal matter, nothing material has yet followed from it. No statute, rule amendment, or reported federal decision has created such a privilege.¹⁴⁵ Altman's proposal therefore remains a policy suggestion, not a legal shield.

There is also a conceptual wrinkle in how the proposal is usually framed. **A privilege ordinarily belongs to the person who confided, not to the custodian of the record:** it is the client's to assert, not the law firm's; the patient's, not the hospital's. An "AI privilege" announced by the chief executive of a provider is therefore *an unusual thing*, a privilege that would attach to records the provider itself holds. That is not incoherent, but **it sits awkwardly with the structure of existing privileges** and helps explain why nothing has followed automatically from the call.

A further objection is normative rather than a logical bar. A privilege of the traditional kind rests on more than a promise not to disclose to outsiders; it presupposes disciplined internal handling by the recipient, the professional secrecy and fiduciary constraint that surround a lawyer or a doctor. It is, in my view, hard to argue persuasively that third parties should be shut out of sensitive chatbot exchanges while the provider reserves broad rights to review, reuse, train on and monetise the same material by default. One can take the other side: a user contracts for a service on defined terms and might claim protection against outsiders whatever rights the provider reserves for its own purposes, much as a tenant's privacy against the world does not dissolve because the landlord keeps a key. But the more durable version of any future privilege is, I believe, one in which external protection and constrained internal handling travel together, which is why a credible Sealed Mode architecture matters here. It would not create privilege by fiat, but it could supply a far more persuasive normative and evidentiary basis for treating some categories of interaction as deserving heightened protection against discovery, breach, or reuse.

3.3.3. *The second question: does using an AI tool forfeit the user's own protection? (Heppner and Warner)*

This second question is a different animal, and it is worth saying how. Neither *Heppner* nor *Warner* is about a provider being compelled to disclose a user's conversations. In *Heppner* the documents were seized from the defendant's own devices under a search warrant in a criminal case; in *Warner* they were materials a self-represented litigant had prepared and was asked to produce in civil discovery. The protection at stake in each is the user's to claim, not the provider's, and the question is not whether the provider's records are reachable but **whether the litigant's own preparation stays protected once it has passed through a consumer AI tool**. The provider is largely a bystander here: it is not the provider's protection to assert, and not something a provider would usually even know to raise.

The reason the question belongs in a chapter on chatbot logs as litigation evidence is that it is **the second route by which the confidentiality users assume fails to become legal protection**. The first route reaches the conversations the provider holds; this one reaches the copies the user keeps, and it decides whether a person can use AI to prepare for a proceeding without forfeiting the protection their own notes and drafts would otherwise enjoy. Two doctrines bear on this, and they work in opposite ways, which is why the cases repay being kept apart. **Work-product protection** can attach to a party's own materials prepared in anticipation of litigation, whether or not a lawyer was involved, so the question for it is whether routing those materials through an AI platform is the kind of disclosure that defeats the

¹⁴⁴ Sarah Perez, "[Sam Altman warns there's no legal confidentiality when using ChatGPT as a therapist](#)" (25 July 2025), TechCrunch. See also [Transcript of This Past Weekend w/ Theo Von, episode #599 with Sam Altman](#) (23 July 2025).

¹⁴⁵ Further research and public debate on how to craft any such privilege would be valuable. As Peter Swire has observed, a federal statute creating an AI-specific privilege would preempt state discovery rules and would raise significant federalism questions under US law, since state courts apply their own rules of evidence and privilege. Any legislative proposal would need to address not only the scope of the privilege but also its relationship to existing state court procedures.

protection. **Attorney-client privilege** works the other way around: it does not attach to a person's solo use of a tool that is not their lawyer in the first place, so the question is not whether AI use forfeits a privilege but whether such use could ever be privileged at all, which on current doctrine it can be only where counsel directs it.

Two federal decisions issued within one week of each other in February 2026 illustrate how unsettled the law remains.

In *United States v. Heppner* (S.D.N.Y., 17 February 2026), Judge Jed Rakoff ruled that 31 documents a criminal defendant had generated using the consumer version of Anthropic's Claude were protected by neither attorney-client privilege nor the work-product doctrine.¹⁴⁶ The facts are instructive. Bradley Heppner, a senior executive indicted on federal securities fraud charges, used Claude on his own initiative to analyse his legal exposure and prepare reports outlining defence strategies. He input into Claude information he had learned from his lawyers and subsequently shared Claude's outputs with his defence counsel, who found them useful and incorporated them into the defence strategy going forward. When the FBI executed a search warrant at Heppner's home upon his arrest, agents seized electronic devices containing approximately 31 documents memorialising these exchanges with Claude.¹⁴⁷

The court held that the documents failed on at least two of the three elements of attorney-client privilege: Claude is not an attorney, the communications were not confidential (given Anthropic's privacy policy permitting data retention and potential third-party disclosure), and Heppner did not use Claude at the direction of counsel. The court also rejected work-product protection because the documents had not been prepared by or at the behest of counsel and did not reflect counsel's strategy at the time they were created. A non-privileged communication, the court held, does not become privileged merely because it is later shown or transferred to counsel.

Heppner is a trial-court decision and has not yet been tested on appeal. It applies traditional privilege doctrine to a new factual setting (a non-lawyer client's solo use of a consumer-tier AI tool, on his own initiative, under terms of service permitting third-party disclosure and training use), and several commentators have observed that different facts may produce different outcomes. Judge Rakoff himself noted that "had counsel directed Heppner to use Claude, Claude might arguably be said to have functioned in a manner akin to a highly trained professional who may act as a lawyer's agent within the protection of the attorney-client privilege", invoking the Kovel doctrine that extends privilege to certain non-lawyer agents of counsel. The Harvard Law Review case note has also argued that the opinion's privilege reasoning veers toward categorical exclusion of client AI use and that a more fact-dependent analysis is preferable. The doctrinal stability of the Heppner approach is therefore an open question that further US litigation, and ultimately appellate review, will need to resolve.

One week earlier, in *Warner v. Gilbarco* (E.D. Mich., 10 February 2026), a different federal court reached the opposite conclusion on the work-product question.¹⁴⁸ The court denied a motion to compel the production of documents prepared by a self-represented litigant using an AI chatbot, holding that they were protected work product and that using ChatGPT did not constitute waiver because "the AI tool (and other generative AI programs) are tools, not persons, even if they may have administrators somewhere in the background". The court emphasised that work-product waiver requires disclosure **"to an adversary or in a way likely to get in an adversary's hand"** and that submitting material to an AI platform did not meet that threshold. The court explicitly rejected the defendant's argument that the mere use of a public AI tool constituted waiver, characterising the motion to compel as seeking "intrusive post-discovery production based on speculation about what might exist in Plaintiff's internal drafting process".

¹⁴⁶ *United States v. Heppner*, No. 25 Cr. 503 (JSR), [Memorandum](#) (S.D.N.Y. 17 February 2026).

¹⁴⁷ The factual summary is drawn from *ibid.* For detailed analyses, see Harvard Law Review Blog, "[United States v. Heppner](#)," March 2026; Proskauer Rose LLP, "[SDNY Addresses Privilege and Work Product Implications of Using Unsecured Public AI Tools](#)," 18 February 2026; New York State Bar Association, "[Loose AI Prompts Sink Ships: How Heppner Shook the Legal Community](#)," 10 March 2026.

¹⁴⁸ *Warner v. Gilbarco, Inc.*, No. 2:24-cv-12333, 2026 U.S. Dist. LEXIS 27355 (E.D. Mich. Feb. 10, 2026). For analysis of the Heppner/Warner split, see White & Case, "[Attorney-client privilege and work product in the age of generative AI](#)," April 2026; Mayer Brown, "[M&A Discovery in the AI Era: Generative AI Communications and Outputs May Become Litigation Ammunition](#)," 27 March 2026; Hogan Lovells, "[The Emerging Rules of the Road Governing AI Prompts and Outputs in Discovery](#)," 23 February 2026; K&L Gates, "[Litigation Minute: Generative AI Data, Attorney-Client Privilege, and the Work-Product Doctrine](#)," 23 February 2026.

The analytical clarity of the Warner approach lies in its insistence on treating waiver, privilege, and work product as distinct inquiries. Even if attorney-client privilege does not attach because the AI is not an attorney, work-product protection may still survive if the AI is treated as a tool used to prepare for litigation rather than as an adversary or ordinary third-party recipient. Heppner, by contrast, placed much greater weight on the third-party character of the AI platform and the provider's privacy policy. Which approach prevails will determine whether users who turn to AI to prepare for legal proceedings can do so without creating a discoverable record of their litigation thinking.

The *Heppner*/*Warner* split thus reveals a fundamental doctrinal uncertainty. Is a consumer AI platform a "tool" (like a word processor, a legal database, or a cloud storage service, in which case work-product protection should survive) or a "third party" (like a friend or business associate, in which case confidentiality may be destroyed)? Judge Rakoff treated Claude as the latter. The *Warner* court treated ChatGPT as the former. Both positions have textual support, and neither is obviously correct.

3.3.4. Why the Heppner reasoning deserves critical scrutiny

Heppner is among the first federal rulings squarely addressing generative AI, attorney-client privilege, and work-product protection, and it will likely be widely cited. But the decision has attracted significant and well-reasoned criticism, and several features of the reasoning deserve scrutiny.

First, *the cloud storage analogy*.¹⁴⁹ As the Harvard Law Review has observed, when clients use Google Docs to draft documents, Gmail to send emails,¹⁵⁰ or iCloud to store materials that are later shared with counsel, courts do not ordinarily ask whether the use of those intermediary cloud services defeats privilege.¹⁵¹ All of those services store user data, all reserve rights to process it under their terms of service, and all could in principle be compelled to disclose it. Yet no court has held that drafting a privileged memorandum in Google Docs destroys the privilege simply because Google's terms of service permit data retention. The *Heppner* court, by contrast, treated the fact that Anthropic's privacy policy permits data collection and potential third-party disclosure as sufficient to destroy confidentiality. The Harvard Law Review concluded that "the *Heppner* court assumed *sub silentio* that Claude was more like a non-attorney human than a tool. One might reasonably question that assumption".¹⁵²

Second, *judicial overreach*. Two detailed critical analyses, published on the same day, converge on the conclusion that Rakoff went further than the case required, but they identify different problems in the reasoning.

In Lawfare, Justin Curl and Mihir Kshirsagar argue that "Judge Rakoff's reasoning is problematic and goes beyond what was needed to resolve the case".¹⁵³ The attorney-client privilege analysis was sufficient on its own: Claude is not an attorney, Heppner did not use it at the direction of counsel, and no recognised privilege attaches on that basis. But Rakoff went on to address confidentiality by treating the chatbot as a third party to whom Heppner had "disclosed" secrets. Curl and Kshirsagar observe that the court's framing becomes inconsistent across the two inquiries: when Rakoff addresses the professional-relationship element, he treats Claude as a software tool; when he turns to confidentiality, the framing

¹⁴⁹ When a user drafts a memorandum in Google Docs, the document is stored on Google's servers, but Google does not read, interpret, or respond to it. The user's relationship with Google Docs is functionally identical to storing a document in a filing cabinet: the cabinet holds the document but does not participate in its creation. A chatbot, by contrast, does not merely store the user's input. It processes it, generates a substantive response, and in some cases retains a structured memory of the exchange. The question is whether this active participation in the generation of content transforms the chatbot from a passive storage medium (like Google Docs) into something more like a third-party interlocutor (like a friend or consultant) to whom a disclosure has been made.

¹⁵⁰ When lawyers began using Internet email in the 1990s, some worried that routing confidential material through a third-party service might forfeit privilege; the bar and the courts soon settled that it did not, treating a reasonable expectation of confidentiality as surviving the use of an ordinary intermediary whose terms permitted access.

¹⁵¹ On the treatment of cloud-based legal workflows under privilege doctrine, see the cases cited in Harvard Law Review Blog, "[United States v. Heppner](#)", March 2026, discussing inter alia the routine treatment of Google Docs, Gmail, iCloud, and Dropbox as non-destructive of privilege when used to prepare or transmit privileged materials. See also Bowditch & Dewey, "[New Tech Meets Old Law: Further Clarification on the Discoverability of AI Output](#)", 7 April 2026, noting that "most off-the-shelf AI programs specifically advise users that prompts and output are retained, are used for LLM training purposes, and may be publicly available", but that this does not resolve the question of whether the use of such programs constitutes a disclosure to a third party sufficient to destroy privilege.

¹⁵² Harvard Law Review Blog, "[United States v. Heppner](#)", March 2026.

¹⁵³ Justin Curl, Mihir Kshirsagar, "[AI and Privilege After United States v. Heppner](#)", Lawfare, 30 March 2026.

becomes more interpersonal, as if Claude were a human interlocutor to whom a disclosure was made. Their central concern is that this reasoning “threatens the access to legal services that AI tools can provide”, penalising precisely the kind of self-directed legal preparation that AI was supposed to democratise.

Separately, David Kemp’s analysis in *Justia Verdict* reaches the same bottom line through a different route: *Heppner* “reached the right result on solid grounds, then bolted on a confidentiality analysis that was unnecessary, incompletely reasoned, and affirmatively wrong in its method”.¹⁵⁴ Kemp’s critique is methodological. He argues that Rakoff treated Anthropic’s broadest contractual reserved rights as dispositive of the confidentiality question without examining “the specific terms, product tier, or training preferences that actually governed Heppner’s use”. Privilege law, Kemp observes, has historically asked whether a party could reasonably expect confidentiality based on the totality of the circumstances, not whether a provider’s boilerplate reserves some theoretical right of access. Kemp warns that it is the overbroad confidentiality reasoning, rather than the uncontroversial holdings on professional relationship and purpose of obtaining legal advice, that future courts will most likely cite, “with potentially damaging consequences for privilege claims in any professional context involving third-party platforms”.

Third, ***the Kovel opening and the access-to-justice problem***. Judge Rakoff left one significant doctrinal opening. He noted that had Heppner used Claude at the direction of counsel, the AI tool “might arguably be said to have functioned in a manner akin to a highly trained professional who may act as a lawyer’s agent within the protection of the attorney-client privilege”, a reference to the *Kovel* doctrine under which communications with non-lawyer professionals retained by counsel are privileged when the professional functions as a conduit for the attorney-client relationship.¹⁵⁵ This dicta is important because it suggests that the same AI tool, used in the same way, could be either privileged or unprivileged depending on whether counsel directed the user to use it. That distinction may be doctrinally sound, but it creates a **paradox**: a sophisticated client with well-resourced lawyers who instruct them to use AI may retain privilege, while an unsophisticated client who turns to AI on their own initiative to prepare for a meeting with counsel loses it. As the Lawfare analysis observed, this outcome is troubling because it penalises precisely the kind of self-help that AI was supposed to democratise.

Fourth, ***the terms-of-service problem***. Judge Rakoff’s confidentiality analysis rested heavily on the fact that Anthropic’s privacy policy permits data collection, potential use for model training, and possible disclosure to third parties including government authorities. But terms of service are drafted unilaterally by the provider, updated frequently, and rarely read by users. If the confidentiality of a client’s legal preparation can be destroyed by a clause in a privacy policy that the client almost certainly never read, then the practical scope of work-product protection in the AI era will be determined not by courts or legislatures but by the drafting choices of technology companies.¹⁵⁶ As Bloomberg Law has noted, the government in *Heppner* “quoted from Anthropic’s consumer privacy policy without presenting documentary evidence establishing which terms actually governed this particular defendant’s account.” The patchwork of privacy terms across AI providers, each with distinct consumer and enterprise tiers updated at different intervals, creates a situation in which the privilege status of a user’s legal preparation may depend on which version of which provider’s terms of service was in effect on the date the prompt was submitted. That is not a stable foundation for a doctrine as fundamental as work-product protection.¹⁵⁷

¹⁵⁴ David S. Kemp, “[The First Federal AI Privilege Ruling Gets the Right Result for the Wrong Reasons](#)”, *Justia Verdict*, 30 March 2026. Kemp also notes that ABA Formal Opinion 512 (2024) identifies whether an AI tool trains on user inputs as the central question for confidentiality, an inquiry that “is exactly the inquiry *Heppner* skipped”.

¹⁵⁵ *United States v. Kovel*, 296 F.2d 918 (2d Cir. 1961). On the *Kovel* opening in *Heppner*, see Lawfare, *op. cit.*; Cf also Bloomberg Law, “[Heppner Shows Attorney-Client Privilege’s Fragility in AI Era](#),” 10 March 2026, observing that Rakoff’s approach “applies even to enterprise-grade AI services used by nonlawyers” and “creates a privilege gap for separately represented executives.”

¹⁵⁶ Bloomberg Law, *ibid.*, noting that “the government quoted from Anthropic’s consumer privacy policy without presenting documentary evidence establishing which terms actually governed this particular defendant’s account.” See also ContractKen, “[AI and the Loss of Privilege: US v Heppner](#),” March 2026, arguing that “SOC 2 tells you the server is secure. It doesn’t tell you whether sending text to that server constitutes a disclosure”.

¹⁵⁷ The legal profession’s own ethics guidance treats careful use of generative AI as compatible with the duty of confidentiality rather than as inherently defeating it. See ABA Standing Committee on Ethics and Professional Responsibility, “[Formal Opinion 512: Generative Artificial Intelligence Tools](#)” (29 July 2024), on a lawyer’s obligations under Model Rule 1.6 when inputting client

Fifth, ***the enterprise/consumer distinction may not hold***. Bloomberg Law has further observed that even the enterprise/consumer distinction may not fully resolve the problem. Rakoff's analysis of the "professional relationship" and "purpose of obtaining legal advice" elements applies regardless of whether the platform is consumer or enterprise grade. A non-lawyer who uses an enterprise AI tool to research legal questions still faces the same doctrinal gap: the AI is not an attorney, and the user is not acting at the direction of counsel. The enterprise tier may solve the confidentiality problem (if the terms of service are sufficiently restrictive), but it does not solve the professional-relationship problem or the purpose-of-obtaining-legal-advice problem. A comprehensive solution would require not merely better terms of service but a new doctrinal framework.

A similar concern has now surfaced in England and Wales. In *UK v Secretary of State for the Home Department (AI hallucinations; supervision; Hamid)* [2026] UKUT 81 (IAC), a case primarily concerned with AI-generated false authorities and professional supervision, the Upper Tribunal also expressed the view that uploading confidential client material to an open or public AI tool could amount to placing it in the public domain, thereby breaching confidentiality and risking waiver of legal professional privilege. The observation was not made in the context of mass chatbot-log discovery, but it confirms that the Heppner question is not limited to U.S. courts: common-law systems are beginning to ask whether use of public AI tools is compatible with the confidentiality on which privilege depends.¹⁵⁸

3.3.5. Conclusion: both questions turn on architecture

Both questions remain unsettled, in different ways. On the first there is simply no privilege: provider-held consumer conversations are reachable in discovery, and Altman's proposed "AI privilege" is an aspiration no statute, rule or reported decision has adopted. On the second the law is not absent but divided: Heppner and Warner disagree on whether a consumer AI platform is a tool, whose use leaves a litigant's own work-product intact, or a third party, to whom disclosure undermines that protection, and the split is unlikely to be resolved at the appellate level for some years. As of the end of April 2026, *Heppner* has not been appealed on the privilege and work-product holdings, and the defendant, who pleaded not guilty, remains before the court; *Warner* was a discovery ruling unlikely to generate an interlocutory appeal. In the interim, users of consumer AI platforms should assume their conversations are neither privileged in the hands of the provider nor reliably protected in their own, while recognising that the foundation for the second assumption is far less secure than *Heppner* standing alone would suggest. What is striking is that both questions ultimately turn on the same architectural fact: whether the provider can read the conversation at all.

The debate could, however, be revived in a more serious and credible way if providers built genuinely sealed modes for especially sensitive categories of conversations. A properly designed Sealed Mode¹⁵⁹ would not itself create privilege by judicial fiat. But it could materially strengthen the factual and normative case for heightened protection by showing that the provider itself has verifiably narrowed internal access, reuse, monetisation, and retention from the moment of collection onward. Under the *Warner* logic (AI as tool), such architectural constraints would reinforce the argument that no disclosure to a third party occurred. Under the *Heppner* logic (AI as third party), they would at least strengthen the confidentiality element by demonstrating that the provider's terms of service and technical architecture are designed to protect, not exploit, the user's input. The first mass-market deployment of an architecture of this kind in the consumer chatbot context has now occurred. Meta's announcement on 13 May 2026 of *Incognito Chat with Meta AI*, which routes conversations between users and Meta AI inside WhatsApp through a Trusted Execution Environment,¹⁶⁰ is a TEE-based design of precisely the kind that the critiques of Heppner and the *Warner* analysis would treat as relevant to the confidentiality assessment. Whether this deployment will, in the medium term,

information into such tools. The Opinion addresses lawyers' use, not a non-lawyer's solo use of a consumer chatbot, but it reinforces that a provider's reserved access rights do not, without more, defeat the confidentiality on which privilege and work-product depend.

¹⁵⁸ Pillsbury Winthrop Shaw Pittman LLP, "[AI and Privilege: Lessons from U.S. v. Heppner and UK Position](#)", 18 March 2026, discussing *UK v. Secretary of State for the Home Department* [2026] UKUT 81 (IAC) and *Ayinde v. London Borough of Haringey*.

¹⁵⁹ See [part I](#) of my study.

¹⁶⁰ Theodore Christakis and Peter Swire, "[The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI](#)", AI-Regulation Papers 2026-05-4, May 2026.

strengthen the doctrinal and normative case for a narrowly defined AI privilege for verified sealed-mode conversations is one of the questions taken up in the conclusions of this study.

3.4. Why de-identification could not fully solve the problem

The courts were willing to place significant weight on de-identification and protective-order controls. But even in the most charitable reading, those measures mitigate risk; they do not erase it.

European data protection law offers a useful conceptual lens for understanding why de-identification of chatbot logs is harder than for conventional datasets. Under the GDPR, personal data that has been pseudonymised but can still be used to re-identify a person remains personal data; only data that has been irreversibly anonymised falls outside the regulation's scope. The EDPB's 2025 Guidelines on Pseudonymisation stress that pseudonymised data remains personal data, that effective pseudonymisation depends on the technical and organisational separation of additional information from the pseudonymised dataset, and that the assessment must account for the means reasonably likely to be used for re-identification.¹⁶¹ The precise boundary between pseudonymised and anonymous data is the subject of significant ongoing European debate, particularly following recent case law and legislative discussions, but that debate does not need to be resolved here.¹⁶² The core point for present purposes is narrower and more practical: chatbot conversations are exceptionally resistant to effective de-identification because the identifying information is woven into the substance of the exchange itself, not confined to detachable metadata fields. Removing a user's name, email address, and phone number from a conversation in which they describe a pending acquisition in a specific industry, a medical diagnosis received last Tuesday from a specific hospital, a legal strategy discussed with their lawyer in a named city, or a family conflict involving details that other people could recognise does not make the conversation anonymous. It makes it a puzzle that anyone with access to the right context, including opposing counsel, expert witnesses, journalists, or determined third parties, could potentially reassemble.

This is precisely the point OpenAI made in its December 2025 memorandum: removing identifiers from logs discussing "details of their finances, immigration status, trading strategies, or investigative journalism" does not make those records "no longer private, sensitive, confidential, or privileged".¹⁶³

The district court nevertheless accepted that existing safeguards were enough for discovery purposes. That is a legal conclusion about proportionality. It is not a finding that confidentiality risk disappeared.

3.5. Why protective orders and judicial safeguards do not eliminate confidentiality risk

A balanced analysis must acknowledge that courts do not compel this kind of discovery casually. The OpenAI orders relied on several safeguards: limiting the production to a sample, de-identifying the data, and placing the material under a protective order. OpenAI itself later said that the data had undergone de-identification and was being made accessible under strict controls designed to prevent copying and printing of non-relevant material.

Those measures can significantly reduce risk. But the historical record shows that they do not eliminate it. At least six distinct residual risks remain.

a) Expansion of the human access circle. The first risk is simple: more people can see the data. OpenAI warned in November 2025 that the Times' request would force it to hand user data to "third parties,

¹⁶¹ European Data Protection Board, [Guidelines 01/2025 on Pseudonymisation](#) (16 January 2025).

¹⁶² See also CJEU, Case C-582/14, *Breyer*, holding that dynamic IP addresses may constitute personal data for a website operator where the operator has legal means reasonably likely to be used to identify the data subject with the assistance of a third party. The more recent EU case law on pseudonymised data, including the litigation concerning the Single Resolution Board and the EDPS, further illustrates that identifiability is contextual and depends on the means reasonably likely to be used by the party receiving or processing the data. For present purposes, the important point is not to resolve the outer boundary of anonymisation under EU law, but to stress that chatbot logs often contain identifying substance within the narrative itself.

¹⁶³ OpenAI, [Memorandum of Law](#), filed 10 December 2025, ECF No. 935, *In re OpenAI, Inc. Copyright Infringement Litigation*, at 2, 10.

including the Times' lawyers and paid consultants".¹⁶⁴ Confidentiality is not binary. It degrades as the number of authorised viewers grows.

b) De-identification leaves substantive secrecy exposed. Even where direct identifiers are removed, the content may still disclose strategic, medical, financial, relational, or psychological secrets. As explained earlier, de-identification is not the same thing as anonymisation, and even pseudonymised data remains capable of revealing a person or at least deeply sensitive substance through context.

c) Misuse or onward disclosure despite protective orders. This is not speculative. A federal judge sanctioned Latham & Watkins for violating a protective order by disclosing to its client an expert report containing attorneys'-eyes-only information in a dispute involving Muddy Waters. The client then shared the report with an employee in an effort to generate negative media coverage.¹⁶⁵ Discovery restrictions are only as strong as the compliance culture of those who receive the data.

d) Spillover into other proceedings or public dockets. A lawyer in Uber sexual-assault litigation was sanctioned for violating a protective order by sharing confidential Uber policy material with other attorneys and using it in unrelated lawsuits. Some of the material was at least temporarily visible on a New Jersey public docket.¹⁶⁶ Even though the judge found limited evidence of substantial public harm, the case shows that court-protected information can move beyond the immediate litigation team and can briefly enter the public record.

e) Breach of law firms and e-discovery vendors. Once sensitive data is produced, it often resides not only with the producing party but also with law firms, experts, e-discovery platforms, cloud repositories, review databases, and managed-service vendors. An e-discovery vendor used by Quinn Emanuel suffered a ransomware attack that may have exposed client information.¹⁶⁷ Jones Day suffered a phishing-linked breach in which hackers accessed client files and posted stolen data online.¹⁶⁸ These incidents are not AI-specific, and that is precisely the point. They show that even if every judge and lawyer acts conscientiously, confidentiality can still fail because the legal-data ecosystem is itself a target.

f) Persistence and function creep. Discovery does not merely expose data to present review. It often creates new copies, derivative work product, notes, expert analyses, excerpts, and machine-readable subsets. Some of these may remain under seal or under protective obligations, but each derivative layer creates opportunities for retention beyond the user's expectation and for future attempts at reuse. Critically, once derivative work product (expert reports, statistical analyses) is created from the chatbot logs, those derivatives may survive even after the underlying logs are returned or destroyed at the end of litigation.

* * *

The most defensible assessment of all these risks is neither alarmist nor complacent. It would be wrong to say that court-ordered safeguards are not important. A sample limited from a much larger universe, de-identification, audited internal handling, secure review environments, and protective orders all reduce the probability of indiscriminate exposure. Courts are entitled to consider those safeguards seriously, and in the OpenAI litigation both the magistrate judge and the district judge did so.

But it would be equally wrong to say that these measures preserve confidentiality against *any* meaningful risk. They do not. Once millions of conversations are preserved under legal hold and made potentially accessible to opposing counsel and experts, confidentiality has already been materially weakened. The historical record shows that protective orders can be violated, confidential material can spill into other proceedings, and law firms or litigation vendors can themselves be breached. None of those precedents perfectly mirrors the disclosure of de-identified AI-chat logs. But the broader litigation record is sufficient to establish that confidentiality in discovery is not fail-safe.

A further development confirms the growing judicial attention to AI-specific risks in the discovery process. In *Morgan v. V2X* (D. Kan., 2026) and *Jeffries v. Harcross Chemicals* (D. Kan., 2026), courts addressed for the first time whether protective orders should restrict the use of AI tools in reviewing

¹⁶⁴ OpenAI, "[Fighting the New York Times' invasion of user privacy](#)" (12 November 2025).

¹⁶⁵ Reuters, "[Column: Latham socked with sanctions in bitter fight involving short seller Muddy Waters](#)" (30 October 2024).

¹⁶⁶ Reuters, "[Uber wins sanctions against lawyer for sexual-assault plaintiffs](#)" (17 February 2026).

¹⁶⁷ Reuters, "[Quinn Emanuel reports cyber attack involving limited client data](#)" (24 July 2023).

¹⁶⁸ Reuters, "[Hackers leak data stolen from Jones Day in latest law firm breach](#)" (6 April 2026).

discovery materials.¹⁶⁹ In *Jeffries*, the court granted a motion to amend the protective order to prohibit the use of publicly accessible AI tools for all discovery materials, not only confidential ones, reasoning that submitting discovery materials to a public AI platform creates risks including the “practical inability to claw back or delete data once it has been incorporated into a model”. The distinction the court drew between “open” AI systems (which may retain submitted data for training) and “closed” or “secure” systems (which operate in a more controlled environment) echoes the Sealed Mode framework proposed in this study: the level of protection depends not only on the legal rules governing disclosure but also on the technical architecture of the tools used to process the disclosed material.

3.6. Could a NYT-style discovery order happen in Europe?

The discovery battles documented in this chapter are products of US civil procedure, and specifically of the broad pre-trial discovery regime under Federal Rules of Civil Procedure 26 through 37, which permits parties to obtain any non-privileged material relevant to a claim or defence, including from third parties. Although the NYT is a US case under US procedural rules, the conversations at stake belong to users worldwide. A European, Asian, or Latin American user whose ChatGPT exchange is swept into a US discovery order has no standing in the US proceeding and may not even know their conversation has been preserved, produced, or reviewed. As noted above, OpenAI’s own FAQ on the 20-million-conversation production order treated GDPR and European privacy rights as relevant concerns, rather than stating that European users were categorically outside the production dispute. The comparative question addressed here is therefore distinct but connected: not only whether US discovery can affect European users, but also whether European procedural and regulatory mechanisms could themselves generate demands for chatbot conversations.

That second question has two halves, and Europe answers them very differently. A chatbot conversation can be sought from the provider, as in the New York Times case, or from the user, as part of an ordinary dispute in which the user is a party. The first route, production by the provider, is hard to replicate in Europe. The second, production by a litigant of their own conversations, is not hard at all, and because the first route is so much weaker here, it is the second that marks the realistic European exposure. Each route is taken in turn, beginning with the provider.

Europe has no identical mechanism. The question, nonetheless, is whether courts or authorities in European jurisdictions, including England and Wales outside the EU and continental Member States within the EU, could produce functionally comparable access to chatbot conversations through narrower procedural or regulatory tools.

No equivalent general discovery regime exists in continental European civil procedure, where the principle is that each party bears the burden of producing the evidence it intends to rely on, and where courts do not ordinarily order one party (much less a non-party) to open its files for adversarial inspection. A US-style order compelling a chatbot provider to produce 20 million or 88 million de-identified consumer conversations to opposing counsel in a copyright case would, as a matter of procedural structure, be extremely difficult to obtain in France, Germany, or most other continental jurisdictions.

But the question is more nuanced than a simple US/Europe contrast suggests. At least four mechanisms could produce functionally comparable results, even if none would reach the scale of the NYT litigation.

England and Wales. The most direct European analogue is the disclosure regime under the Civil Procedure Rules (CPR) in England and Wales. CPR Part 31 requires parties to disclose documents relevant to the issues in the case, including documents adverse to their own position. Critically, CPR 31.17 empowers courts to order disclosure by non-parties where the documents sought are likely to support the applicant’s case or adversely affect the case of another party, and where disclosure is necessary to dispose fairly of the claim or to save costs.¹⁷⁰ English courts have shown increasing willingness to order disclosure of digital evidence, including material held on personal devices and

¹⁶⁹ See Sidley Austin LLP, “[Generative AI in Discovery: Protective Orders as an Emerging Point of Dispute](#)”, 6 April 2026.

¹⁷⁰ Civil Procedure Rules (England and Wales), Part 31, [Rule 31.17](#). The court may order non-party disclosure only where the documents sought are “likely to support the case of the applicant or adversely affect the case of one of the other parties to the proceedings” and disclosure is “necessary in order to dispose fairly of the claim or to save costs”.

informal messaging platforms such as WhatsApp, recognising that “work is carried out on personal electronic devices” and that “any mingling of personal and work data will not itself be sufficient to circumvent a disclosure order”.¹⁷¹ While no English court has yet ordered bulk disclosure of consumer chatbot conversations, the jurisdictional tools to do so exist. A copyright claimant suing an AI provider in the English courts could, in principle, seek non-party disclosure of chatbot conversation logs under CPR 31.17, and the court would assess the application under the same relevance and necessity criteria that apply to other categories of electronic evidence. The scale would likely be narrower than in the US, but the mechanism is available.

The EU IP Enforcement Directive. In continental Europe, the most relevant instrument for copyright-related disputes is the IP Enforcement Directive (2004/48/EC), which requires Member States to provide judicial authorities with the power to order evidence preservation and production in intellectual property proceedings.¹⁷² Article 6 requires Member States to ensure that courts may order the opposing party or a third party to present evidence that lies within their control, subject to confidentiality protections. Article 7 empowers courts to order “prompt and effective provisional measures to preserve relevant evidence in respect of the alleged infringement”, including “the physical seizure of the infringing goods, and, in appropriate cases, the materials and implements used in the production and/or distribution of these goods and the documents relating thereto”. Article 8, the “right of information”, may also become relevant where claimants seek information on the origin, distribution channels, scale, or commercial use of allegedly infringing activity. Although Article 8 is not a general discovery provision, it reinforces the point that EU IP enforcement law contains targeted mechanisms for compelling information in infringement proceedings.

These provisions were explicitly modelled on common-law discovery mechanisms and have been described as introducing “evidence-related provisions and the Right to Information [that] are similar to discovery rules in the United States”.¹⁷³ In a case analogous to *NYT v. OpenAI*, a copyright holder in a Member State could invoke Article 6 to request the production of chatbot output logs that demonstrate reproduction of protected works, and Article 7 to seek preservation of such logs pending resolution of the case. The scope would be controlled by the principle of proportionality, and courts would need to balance the claimant’s interest in proving infringement against the privacy interests of uninvolved users. But the mechanism exists, and it has not been tested in the context of consumer AI logs.

National procedural mechanisms. Several national procedural codes provide additional tools. In France, for instance, Article 145 of the Code de procédure civile allows any party who has a “motif légitime” (legitimate reason) to request, even before proceedings on the merits have commenced, that a court order “mesures d’instruction” (investigative measures) that may include the seizure of documents held by a third party, provided the measure is legally admissible and proportionate. The French “saisie-contrefaçon” in intellectual property cases goes further, allowing rights holders to obtain a judicial order authorising a bailiff to enter premises, seize evidence of infringement, and make copies of relevant documents and electronic files.¹⁷⁴

Regulatory investigations. Beyond civil litigation, regulatory proceedings could produce functionally equivalent access to chatbot data. Data protection authorities conducting investigations under GDPR Articles 58(1)(a) and (e) have the power to order the controller to provide all information necessary for the performance of their tasks and to obtain access to any premises of the controller, including to data processing equipment and means. Competition authorities investigating potential abuse of dominance or anti-competitive agreements could, in principle, request access to chatbot interaction data as part of their investigative powers. Under the Digital Services Act, the European Commission has investigative

¹⁷¹ On the English courts’ increasing willingness to order disclosure of digital evidence from third-party devices, including personal mobile phones and informal messaging platforms, see Latham & Watkins, “[English Court Increasingly Willing To Order Disclosure From Third-Party Devices](#)”, 17 February 2021, discussing *Phones 4U Ltd (In Administration) v. EE Ltd & Ors* [2021] EWCA Civ 116 and related cases.

¹⁷² [Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights](#), Articles 6 (evidence), 7 (measures for preserving evidence), and 8 (right of information). See also European Commission, [Guidance on certain aspects of Directive 2004/48/EC, COM\(2017\) 708 final](#), 29 November 2017.

¹⁷³ WilmerHale, “[New EU Directive on IP Enforcement](#)”, 16 September 2004.

¹⁷⁴ Code de procédure civile (France), Article 145: « S’il existe un motif légitime de conserver ou d’établir avant tout procès la preuve de faits dont pourrait dépendre la solution d’un litige, les mesures d’instruction légalement admissibles peuvent être ordonnées à la demande de tout intéressé, sur requête ou en référé ». On the saisie-contrefaçon as a distinctive French evidence-gathering mechanism in IP cases, see Code de la propriété intellectuelle, Articles L. 332-1 et seq. (copyright) and L. 615-5 et seq. (patents).

authority over very large online platforms and very large online search engines (Articles 65-66), including the power to request information and to conduct inspections. If a chatbot provider is designated as a VLOP or VLOSE, these investigative powers could extend to user interaction data. While regulatory investigations differ from adversarial civil discovery in structure and purpose, they share a common feature: the capacity to compel a provider to produce sensitive user data to an external authority.

The GDPR as both constraint and complication. European data protection law adds a layer of complexity that has no direct US equivalent. Any production of personal data for litigation purposes constitutes processing under the GDPR and requires a lawful basis. Article 6(1)(c) (compliance with a legal obligation, such as a court order) and Article 6(1)(f) (legitimate interests) are the most relevant bases, but both are subject to proportionality. For special categories of data under Article 9, including health data, data revealing political opinions, and data concerning a person's sex life, the conditions for lawful processing are more restrictive, and Article 9(2)(f) permits processing only where it is "necessary for the establishment, exercise or defence of legal claims". Chatbot conversations will frequently contain special-category data. A court ordering the production of millions of chatbot conversations to opposing counsel would need to assess not only relevance and proportionality under procedural rules but also compliance with GDPR requirements, including data minimisation (Article 5(1)(c)) and purpose limitation (Article 5(1)(b)). These constraints would almost certainly result in narrower production orders than those issued in the NYT case, and would likely require strong de-identification, access restrictions, and other safeguards. But they would not categorically prevent court-ordered production of chatbot data. The GDPR constrains the scope and conditions of disclosure; it does not create an absolute bar.¹⁷⁵

Demands aimed at the user. Everything to this point concerns whether a European court or authority could compel the provider to surrender conversations. The route most likely to reach an ordinary European runs the other way, through the user's own case, and here the gap with the United States almost closes. European systems do not permit broad pre-trial discovery, but they do let a court reach a party's own documents. In England and Wales, standard disclosure under CPR Part 31 requires each party to disclose the documents within its control that bear on the issues, including those adverse to its own case, so that a litigant's own chatbot conversations are disclosable on the same basis as their emails or messages. In France and Germany, although there is no general discovery, a court may order a party, on a specific and reasoned basis, to produce a document in its possession. Data protection operates here much as the Stored Communications Act does in the United States. A provider cannot lawfully pass one user's conversations to that user's opponent, and an opposing litigant has no freestanding right to demand another person's personal data from the provider, so in an ordinary dispute the lawful and practical route to the content runs through the account holder rather than the company.

Assessment. The two routes deserve different answers. A US-style order compelling the bulk production of tens of millions of consumer conversations from the provider remains, for now, a distinctively American phenomenon: the procedural culture, the breadth of pre-trial discovery, and the absence of a general data protection framework in US federal litigation combine to produce outcomes that would be very difficult to replicate in most European jurisdictions, even though the English disclosure regime, the IP Enforcement Directive, national preservation and seizure mechanisms, and regulatory investigative powers each leave a narrower door open, and the GDPR would tighten that door rather than open it. The route through the user needs no such qualification. A European litigant's own chatbot conversations are reachable in their own case today, through ordinary disclosure and production rules, with data protection channelling the demand towards the user rather than barring it. The spectacular form of the problem, the mass production order, is for now American; the everyday form, developed at the end of the next section, already reaches both sides of the Atlantic. As AI copyright litigation, product-liability claims, and regulatory investigations multiply across the continent, the narrower provider routes will be tested too, but for ordinary users the live question is not whether their

¹⁷⁵ The CJEU has confirmed that the GDPR does not create an absolute bar to the disclosure of personal data in litigation: see, e.g., Case C-275/06, [Promusicae v. Telefónica](#) [2008] ECR I-271, holding that Member States must ensure a fair balance between intellectual property rights and data protection rights when implementing evidence-disclosure obligations. See also Case C-268/21, [Norra Stockholm Bygg](#), EU:C:2023:145, confirming that the GDPR applies to the production of documents containing personal data in civil proceedings and that courts must balance the interests of the party seeking evidence against the data protection rights of the persons concerned.

conversations will be drawn into litigation so much as by which of the two routes, and the answer is mostly the second.

3.7. Beyond the New York Times: from bulk discovery to the everyday litigant

The New York Times dispute is the clearest public example of the mass discovery of chatbot conversations, but it is the exceptional case, not the representative one: the conversations sought belonged to millions of people who were strangers to the suit, and they were demanded from the provider in bulk. Almost every other route by which a conversation becomes evidence is narrower than that, and almost every other route runs closer to home. The previous section showed that the divide between a demand aimed at the provider and one aimed at the user holds on both sides of the Atlantic. This section follows that divide down from the headline case to the point at which it touches an ordinary person: the litigant whose own conversations surface in their own dispute.

The cases fall into the two groups the previous section set apart. In the first, the conversations are produced by the provider, because the AI company is itself a party to the litigation, sued over what its product did, or compelled to hand over the records of users who are strangers to the suit. In the second, the company is a bystander, and it is the user who produces the conversations, a litigant whose own exchanges are drawn into their own dispute. The first group is the more conspicuous; the second is the more common, and the party most exposed in it is not a company defending a claim but the individual who never expected to be in court at all.

3.7.1. Conversations sought from the provider

In each of these cases the AI company is a party, and the conversations surface from its own files:

Copyright and training cases. The closest comparator is Anthropic. In *Concord Music Group v. Anthropic*, a federal court in California ordered production of a random sample of five million Claude prompt-output pairs, drawn under a court-approved protocol from relevant pre-suit and post-suit periods.¹⁷⁶ The same order recognised that Anthropic’s users had a privacy interest in their use of Claude, while still compelling large-scale production under a protective order. That is already enough to show that the OpenAI matter is not a one-off curiosity. In major AI copyright litigation, courts have begun to treat prompts and outputs as ordinary electronically stored information that may be sampled, statistically analysed, and produced at scale.

Safety, wrongful-death, and product-liability cases. Google and Character.AI faced a Florida mother’s lawsuit alleging that Character.AI’s chatbots contributed to her teenage son’s suicide; the matter settled in January 2026.¹⁷⁷ Google’s Gemini is now the subject of a wrongful-death suit in which the plaintiff alleges that the chatbot encouraged dangerous emotional attachment and facilitated a user’s suicide.¹⁷⁸ In disputes like these, the relevant conversations are often the most direct evidence of what the system said, how the user responded, and whether guardrails failed. For now, these are mainly cases of targeted, individualised log discovery rather than bulk disclosure. But they show how central chatbot conversations can quickly become in litigation.

Cases involving minors and exploitative interactions. Filings in New Mexico litigation against Meta included internal employee messages obtained through discovery, allegedly showing resistance within Meta to stronger guardrails for minors exposed to sexualised AI conversations, a characterisation Meta disputes.¹⁷⁹ This is not the same as a bulk handover of ordinary user logs. But it shows that once a

¹⁷⁶ [Concord Music Group, Inc. v. Anthropic PBC, Order re Discovery Disputes](#), No. 24-cv-03811-EKL (SVK) (N.D. Cal. 8 August 2025).

¹⁷⁷ Reuters, “[Google, AI firm must face lawsuit filed by a mother over teen son’s suicide, U.S. court says](#)” (21 May 2025). See also Reuters, “[Google, AI firm settle Florida mother’s lawsuit over son’s suicide linked to chatbot](#)” (7 January 2026).

¹⁷⁸ Reuters, “[Lawsuit says Google’s Gemini AI chatbot drove man to suicide](#)” (4 March 2026).

¹⁷⁹ Reuters, “[Meta CEO Zuckerberg blocked curbs on sex-talking chatbots for minors, court filing alleges](#)” (27 January 2026). Meta responded that the State had been ‘cherry-picking documents to paint a flawed and inaccurate picture’, and that the documents in fact showed its chief executive directing that explicit AI companions not be available to younger users. Ibid.

complaint alleges chatbot-induced harm, the actual exchanges and the surrounding internal records can become central evidence.

Professional-services and unauthorised-practice cases. Nippon Life sued OpenAI in federal court in Chicago, alleging that ChatGPT effectively engaged in the unauthorised practice of law by helping a user reopen a settled disability dispute and generate a large volume of filings, some allegedly relying on fictitious legal authorities.¹⁸⁰ That case too is not a mass-discovery case, at least not yet. But it shows how quickly user interactions can become pivotal evidence in litigation about legal advice, compliance, consumer deception, or professional reliance. Cases of this type could eventually produce discovery requests for the conversations of third parties who relied on ChatGPT's advice, not just the specific user involved.

3.7.2. Conversations sought from the user

In the remaining cases the company stands outside the dispute, and the conversations are the litigant's own:

Corporate governance and M&A disputes. In *Fortis Advisors LLC v. Krafton, Inc.* (Delaware Court of Chancery, 16 March 2026), the court found that Krafton's CEO had consulted ChatGPT while developing strategies to avoid paying a \$250 million earnout owed to the founders of an acquired studio, and had then implemented elements of the AI-assisted strategy. The court quoted the CEO's ChatGPT conversations directly in its opinion as evidence relevant to intent and breach. Critically, the CEO had turned to ChatGPT after receiving legal advice warning against the course of action he was considering. The case is analytically distinct from the NYT bulk-discovery dispute: it involves a litigant's own chatbot conversations being used as direct evidence of intent and strategy, not the mass production of uninvolved users' logs. But the lesson for this chapter is complementary. The NYT case shows that third parties' chatbot conversations may be discoverable at scale. *Krafton* shows that a party's own chatbot conversations can be among the most damaging evidence in the case, precisely because users interact with chatbots in a candid mode they might never adopt in an email to a colleague.¹⁸¹

The everyday litigant. *Krafton* involved a chief executive and a quarter-billion-dollar earnout, but nothing in the reasoning turns on the size of the stakes or the prominence of the party. The same logic reaches the ordinary person, and this is where the risk stops being abstract. Chatbot conversations are electronically stored information, and in civil litigation a party can be required to produce relevant, non-privileged electronically stored information from their own accounts, just as parties are now routinely ordered to produce text messages, emails and social-media activity. A spouse in a contested divorce, an employee bringing a dismissal claim, a parent in a custody dispute, a driver in a personal-injury suit: each may be asked, in discovery, to hand over the conversations they had with a chatbot about the very matters in issue, and each may be under a duty to preserve those conversations once litigation is reasonably anticipated. How far this exposure materialises in any given case turns on the interaction of provider retention periods, the point at which a duty to preserve attaches, and the law of spoliation, which raise technical and legal questions this chapter cannot pursue. The point in the text is the legal one: such conversations are discoverable and carry no privilege. Whether a particular conversation still exists, and who may reach it, is a separate and contingent question.¹⁸²

The result is a quiet gap between how these conversations feel and how they travel. A user confides in a chatbot in the register described in Chapter 1, candid, exploratory, often unguarded, on the tacit assumption that the exchange is theirs. What the discovery rules see is a dated, searchable, first-person record of what that person knew, intended, feared or planned, in their own words and in their own

¹⁸⁰ Reuters, "[OpenAI hit with lawsuit claiming ChatGPT acted as an unlicensed lawyer](#)" (5 March 2026).

¹⁸¹ Delaware Court of Chancery, *Fortis Advisors LLC v. Krafton, Inc.*, opinion of 16 March 2026. See Alston & Bird Privacy Blog, "[Your AI Chats May Be Used Against You: CEO's ChatGPT Records Appear in Judicial Opinion Concerning \\$250 Million Earnout](#)", 28 April 2026.

¹⁸² Deletion is the user's only real shield, and it is a leaky one. It is self-help, not law: the rules give the user no privilege and nothing they may withhold, so a conversation that still exists must be produced. Nor does deleting reliably destroy it, since the provider retains the conversation for a period after the user removes it, and indefinitely under a litigation hold; what the user gains by deleting is the loss of their own access and of the opponent's easy route, not the erasure of the record, which persists at the provider and can resurface wherever the provider becomes reachable. And most users delete nothing at all, because it never occurs to them that a chatbot history might be read in court. Deleting relevant material once litigation is in prospect is, in any event, spoliation.

account, with no privilege to shield it and a duty to preserve it once a dispute is on the horizon. It is the candour that makes the record so useful to an opponent. The mass production in the New York Times case is the dramatic form of the problem; this quieter form is the one far more likely to reach an individual, and it requires being a party not to any headline litigation but only to an ordinary one.

3.8. Why similar discovery fights may persist and multiply

Neither form of the problem is likely to recede. Several forces may sharpen both the provider-side fights and the user-side ones, and turn demands for chatbot conversations into a recurring feature of litigation rather than a curiosity.

First, chatbot conversations are often the best contemporaneous evidence of product behaviour in the wild. Unlike a static software bug, harmful or infringing chatbot outputs are interaction-specific. These conversations can reveal not only what the model said, but also what the user asked, whether the answer was reformulated after repeated prompting, and whether the interaction formed part of a broader pattern.

Second, the social functions of chatbots are expanding. Users do not employ them only for generic drafting. They use them for legal analysis, emotional support, software development, medical inquiry, business planning, and educational work. The more central the chatbot becomes to real-life decision-making, the more likely conversations are to be probative in disputes over causation, reliance, negligence, foreseeability, market substitution, and damages.

Third, the data itself is unusually scalable. A traditional paper archive or even a conventional email corpus imposes practical friction. Centralised chatbot logs are natively machine-searchable, sampleable, and analysable. That makes them especially attractive for plaintiffs seeking statistical proof or evidence of patterns of conduct.

Fourth, as Section 3.3 showed, the law offers no privilege or comparable buffer for ordinary chatbot use, so courts applying ordinary relevance and proportionality principles will continue to face strong arguments for disclosure.

Finally, the “user-side” form multiplies with adoption. The forces above mostly explain why litigation against the AI companies will recur; the quieter fights described in the previous section grow for a simpler reason. As hundreds of millions of people route into these systems the candid thinking they once kept to themselves, each builds up a dated, searchable record that will increasingly be sought not in disputes about AI but in the ordinary ones, a divorce, an employment claim, a business falling-out, another corporate claim... The more central the chatbot becomes to how people decide things, the more often its record will be asked for in their own cases.

3.9. Conclusions and tentative recommendations

The discovery chapter of this study documents **a threat to chatbot privacy** that is less visible than government access or proactive corporate disclosure but no less important, and it arrives by **two distinct routes**.

Along the first, the conversations are sought from the provider: the New York Times litigation has established, through binding judicial orders, that consumer chatbot conversations are discoverable at scale, the *Concord v. Anthropic* order confirms that this reaches more than one provider, and the expansion from twenty to eighty-eight million conversations shows how fast the scope can grow once discoverability is accepted.

Along the second, the conversations are a litigant’s own: *Krafton* shows that a person’s own exchanges with a chatbot can become the most damaging evidence in their case, and that this is the route most likely to reach an ordinary individual, in a divorce, an employment claim, or a business dispute.

The chapter **also separated the two questions the loose phrase “AI privilege” runs together**.

The first, whether any privilege shields a provider's records of consumer conversations, has a clear answer: none does.

The second, whether using a consumer AI tool costs a litigant the work-product protection their own preparation would otherwise carry, does not: *Heppner* held that it did, *Warner* that it did not, and that divergence, which arose in a criminal prosecution and a civil suit respectively, will need to be resolved. Recent protective-order decisions in Kansas add that courts are beginning to treat the submission of discovery material to public AI tools as itself a confidentiality risk.

Beneath both routes lies a single mismatch between how people use consumer chatbots and how the law treats the records that result. Users increasingly speak to these systems as though entering a semi-confidential personal space; the law treats the resulting exchanges as ordinary, non-privileged records, discoverable from the provider when relevance and proportionality are satisfied and producible by the user in their own case like any other document they hold. The conversations enjoy no legal protection of their own. What protection a user has comes instead from architecture and habit, from whether the provider retains the conversation, whether the user deletes it, and whether a confidential pathway was built into the product at all, none of which is a legal right and all of which can fail. That gap between felt confidentiality and the absence of any legal shield will only widen as chatbots move further into health, law, education, finance, and emotional companionship.

On the first route, a further problem is representation. The users whose conversations are preserved, sampled, de-identified, produced, and reviewed are generally not parties to the litigation, receive no individual notice, and have no meaningful opportunity to object. Their privacy interests are asserted, if at all, by the provider, whose incentives may only partially align with theirs. This is a recurring feature of large-scale platform litigation, but chatbot conversations make the problem more acute because the affected material may include health concerns, legal strategies, intimate relationships, credentials, business plans, and private emotional disclosures. Courts confronting bulk chatbot discovery should therefore treat **absent-user privacy** not as an incidental side issue but as a distinct interest requiring structured representation through heightened proportionality analysis, privacy-focused special masters, amicus participation, or other procedural safeguards.

The two questions the chapter separated call for two different answers.

The first, the absence of any privilege over a provider's records, is a question about the conversation as the provider holds it, and here the lever is architecture. Whether a confidentiality claim can ever rest on these conversations depends not on the user's subjective sense of privacy but on objective indicators present when the conversation takes place: the provider's terms, its retention and access practices, and whether it has constrained its own reach. Standard consumer terms that reserve broad rights to retain, train on, and share conversations supply no such indicators, and courts will cite them as proof that confidentiality was never established. A provider that wants its users' conversations to be capable of protection must build that capability into the product, which is what **Sealed Mode** is for.

The second question, whether a litigant's own AI-assisted preparation keeps the work-product protection it would otherwise have, is not about the provider at all. It is about the user's materials in the user's case, and it turns on doctrine: on whether a court treats the chatbot as a tool the litigant used, as *Warner* did, or as a third party to whom the work was handed, as *Heppner* did. Architecture cannot resolve that. Only the courts or the rule-makers can, and until they do, the protection a litigant receives will turn on the accident of which view their judge takes.

There is also a second-order consequence that extends beyond the immediate confidentiality harm. Once it becomes widely known that chatbot conversations can be produced in litigation, including litigation to which the user is not a party, users may begin to self-censor, avoid sensitive topics, or migrate to providers with stronger architectural protections. This chilling effect is analytically distinct from the direct confidentiality risk and may ultimately be more damaging: the value of a chatbot as a tool for reflection, honest self-examination, and creative exploration depends on the user's willingness to engage without inhibition. If that willingness erodes because users learn that their conversations may become evidence, the social utility of the entire technology diminishes. The existence of a Sealed Mode option could partially address this concern: users who know that a protected pathway exists may use it precisely for the conversations they would otherwise suppress.

The European picture, set out in Section 3.6, divides along the same two routes. The first, a bulk order compelling a provider to produce conversations at scale, is for now a distinctively American phenomenon: continental procedure has no general discovery, and the narrower tools that exist, non-party disclosure in England and Wales, the IP Enforcement Directive, national seizure mechanisms, and regulatory investigations, would reach nothing like the scale of the New York Times order, with the GDPR tightening even further the door. The second route needs no such qualification. A European litigant's own chatbot conversations are reachable in their own case today, through ordinary disclosure and production rules, and data protection operates here much as the Stored Communications Act does in the United States, barring the provider from handing the content to an opponent and so channelling the demand to the user. The everyday exposure is therefore not an American peculiarity but a feature on both sides of the Atlantic; only the spectacular version is, for the moment, confined to the United States.

A final note on existing institutional work. The intersection of AI and civil discovery is already the subject of substantial US institutional engagement, much of it predating the New York Times litigation. The Sedona Conference's Working Group 1 on Electronic Document Retention and Production has been examining generative AI and eDiscovery since 2023, and in September 2025 formed a dedicated brainstorming group on the discoverability and admissibility of GenAI-related data, prompts, inputs, and outputs. The Sedona Conference also launched a new Working Group 13 on Artificial Intelligence and the Law at its January 2025 inaugural meeting, with a dual mission of coordinating AI-related activities across the existing Working Groups and developing principles and best practices on AI in other legal contexts.¹⁸³ State bar associations have issued ethics opinions addressing lawyers' use of generative AI, and, among several other developments, the Administrative Office of the US Courts distributed interim AI guidance to federal courts in July 2025.¹⁸⁴ To my knowledge, however, none of this existing work has yet specifically addressed the distinctive situation foregrounded by the New York Times litigation: large-scale discovery of the substantive content of consumer chatbot conversations belonging to millions of non-party users. The recommendations that follow are offered as a contribution to that emerging conversation, from a non-US legal perspective, in the hope of being a useful input into the work the relevant US institutions are already undertaking.

The recommendations below address both the legal and architectural dimensions of this problem.

Tentative Recommendations on Discovery and Evidence

#	Recommendation
9	Design products with discovery exposure as an explicit consideration. The discovery precedent demonstrates that stored chatbot conversations are vulnerable to large-scale production orders, and the volume a provider retains directly shapes the scope of any future order. Providers should treat architectural choices, retention limits, the design of memory features and conversation history, and Sealed Mode pathways for the most sensitive categories, as decisions with discovery implications and not only privacy ones. Reducing what is retained, giving users genuine control over retention and deletion, and architecting a verifiable Sealed Mode pathway for the conversations that warrant the strongest protection narrow the surface available for litigation capture as well as for the other access pathways analysed in this study.
10	Pursue a limited AI privilege for the most sensitive chatbot conversations, supported by architecture rather than by terms of service alone. This recommendation addresses the first of the two questions the chapter separated, the absence of any privilege over a provider's records; the second, work product, is taken up immediately below. No general privilege

¹⁸³ For The Sedona Conference's AI-related activities, see The Sedona Conference, "[The Sedona Conference's AI-Related Activities](#)" (describing WG1's brainstorming group on machine learning and AI in discovery); Doug Austin, "[GenAI-Related Data and The Sedona Conference](#)", eDiscovery Today (10 September 2025) (announcing the WG1 brainstorming group on discoverability and admissibility of GenAI-related data). On Working Group 13, see The Sedona Conference, "[Working Group 13 on Artificial Intelligence and the Law: Inaugural Meeting 2025](#)", 15-17 January 2025, Phoenix, Arizona.

¹⁸⁴ [Letter from Judge Robert J. Conrad](#), Director, Administrative Office of the US Courts, to Sen. Charles Grassley, Chairman, Senate Judiciary Committee (October 2025) (describing interim federal-judiciary AI guidance distributed 31 July 2025).

	protects chatbot conversations from discovery under current US law, and recognising one would require legislative, rule-making, or judicial action. None of these routes is foreclosed, but each requires that the conversations the privilege would protect display objective indicators of confidentiality at the time they take place. Consumer conversations conducted under standard terms that permit retention, training use, and disclosure to third parties cannot supply those indicators; conversations conducted under verified Sealed Mode conditions, where the provider has demonstrably constrained its own access, retention, and reuse from the moment of collection, can. The recommendation is therefore directed first at providers: build the architectural and contractual foundations on which a privilege claim can rest. With those in place, the case for a limited AI privilege, recognised by whichever of the three vehicles is most appropriate, becomes substantially stronger. Any such privilege would retain the crime-fraud exception that accompanies established privileges.
11	Resolve the work-product question raised by <i>Heppner</i> and <i>Warner</i>. The divergence between <i>Heppner</i> , which treated a litigant's AI-assisted materials as disclosed to a third party and so unprotected, and <i>Warner</i> , which treated the chatbot as a tool the litigant used and left work-product protection intact, leaves the protection of a party's own preparation turning on which view a court happens to take. This is a doctrinal question that the privilege proposed above would not reach, because it concerns the user's own materials rather than the provider's records. Courts, and in time the rule-making bodies, should clarify that using a general-purpose AI tool to assist one's own litigation preparation does not by itself waive work-product protection, any more than using a word processor or a search engine does, while leaving the ordinary limits of the doctrine, including the crime-fraud exception, intact. Until that clarity arrives, a litigant should assume that AI-assisted preparation is discoverable.
12	Develop chatbot-specific proportionality guidance for civil litigation. Federal Rule of Civil Procedure 26 requires proportionality, but no stable guidance yet exists on how it should be assessed where the discovery target is the bulk conversational record of millions of non-party users. The substantial US institutional work already under way on generative AI and eDiscovery, notably at The Sedona Conference, would be a natural setting in which to address the situation the New York Times litigation has foregrounded.
13	Strengthen de-identification and govern derivative materials. The courts in the New York Times litigation accepted de-identification as a sufficient safeguard, but de-identification of chatbot conversations is harder than for conventional records, because the substance of a conversation can itself be identifying, as European data protection law has long recognised in distinguishing pseudonymisation from anonymisation. Providers should develop chatbot-specific de-identification protocols that address not only direct identifiers but also rare fact combinations and references to unique personal circumstances. The protective orders that accompany large-scale production should also govern what happens at the end of the litigation, not only to the produced conversations themselves but to derivative materials such as excerpts, expert datasets, summaries, and embeddings, limiting their use to the specific proceeding.
14	Warn users, especially in sensitive contexts, that their conversations may be discoverable, and give them real control over what is kept. Sam Altman's July 2025 acknowledgment that ChatGPT conversations lack legal confidentiality was important but insufficient. Providers should inform users clearly, and in plainly visible form, that ordinary consumer conversations may be subject to lawful production in litigation, and that the exposure runs in two directions: a user's conversations may be swept into litigation to which they are a stranger, and, just as importantly, a user's own conversations may be demanded in their own divorce, employment, or business dispute, where no privilege will shield them. The warning should be especially prominent when users enter health, mental-health, legal, financial, or other high-sensitivity contexts. Because in an ordinary dispute the user's only practical protection is to limit what is

	retained, the warning should be paired not only with a protected pathway such as Sealed Mode but with genuine, easily used controls over retention and deletion.
--	--

4. Cyberthreats, Insider Risks and Other Data Breaches

The previous chapters examined three lawful or quasi-lawful routes by which chatbot logs can leave the confidential space imagined by users: proactive disclosure by the provider, compelled access by the state, and discovery. This chapter examines a fourth route: **exposure through failure**. The relevant failure may be technical, as when a database is left open or a software bug reveals other users' data. It may be organisational, as when employees, contractors, or vendors have broader access than users expect. It may be user-facing, as when an interface turns a private conversation into a public artefact through confusing sharing design. Or it may be adversarial, as when attackers steal credentials, inject malicious prompts, hijack connected tools, or abuse agentic permissions to extract data.

This chapter does not attempt to provide a complete cybersecurity taxonomy. Its purpose is narrower: to show that chatbot confidentiality cannot be assessed only through privacy policies, government-request rules, or discovery doctrine. It must also be assessed through resilience. The more intimate chatbot logs become, the more damaging their accidental or malicious exposure becomes. Retention, memory, operational sharing, connected apps, and agentic capabilities therefore matter twice: first as privacy design choices, and second as determinants of breach impact.

Part 1 of this study showed that consumer chatbot conversations are not held inside a sealed vault. They may be used for training, reviewed by humans, processed by vendors, shared across operational systems, connected to other services, and retained in ways most users do not understand. It also distinguished operational sharing with processors and vendors from a sale of personal data: such sharing is often necessary, often legally constrained, and sometimes privacy-protective. But Part 1 also stressed that users typically cannot know which categories of vendors and internal teams may access plaintext chat content, under what constraints, and for how long.¹⁸⁵ That opacity becomes more consequential once one adds cybersecurity risk. A vendor relationship that is lawful under GDPR Article 28 (concerning data processors) can still create a new attack surface. A support workflow that is operationally necessary can still expand insider access. A memory feature that improves personalisation can also create a richer target. The question is whether the architecture is sufficiently hardened for the sensitivity of the data at stake.

4.1. Why chatbot logs are a distinctive breach target

Not all personal data has the same breach consequences. A leaked email address, password hash, or billing record can be serious. A leaked chatbot log can be serious in a different way: a narrative record of a person's fears, intentions, symptoms, legal strategy, relationship conflict, political doubts, business plans, code, credentials, or attempts to understand a crisis. The difference is not only that chatbot logs may contain sensitive data but also that they often contain sensitive data in context. A single conversation may combine identity, motive, chronology, emotion, and factual detail in a way that makes the content self-identifying even after ordinary identifiers have been removed.

This is why the legal definition of a personal data breach is important but insufficient. Under the GDPR, a personal data breach is a breach of security leading to accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data. Controllers must notify the competent supervisory authority without undue delay and, where feasible, within 72 hours, unless the breach is

¹⁸⁵ Part 1 explains that operational sharing with downstream service providers, such as cloud hosts, safety auditors and customer-support tooling, is legally and functionally different from selling data, but that the practical trust question is whether users can understand and verify who accesses their conversations. See T. Christakis, [You Trust Your Chatbot With Everything. Should You? Part 1](#), AI Regulation Papers, 26-03-2, March 2026.

unlikely to result in a risk to the rights and freedoms of natural persons; communication to affected individuals is required where the breach is likely to result in a high risk.¹⁸⁶ These rules are essential, but chatbot logs test them in an unusual way. The risk assessment cannot be based only on whether names, emails, or account IDs were exposed. It must also ask what the conversations reveal substantively: health status, children, legal vulnerability, sexual life, immigration status, workplace misconduct, psychological distress, trade secrets, or security credentials. A further dimension sits outside the individual-harm calculus altogether. Chatbot logs can also hold material of value to hostile actors, from the operational details a user discloses to the simple fact of who is asking what, which is one reason advanced persistent threats and foreign intelligence services treat large conversational datasets as a prize. This is part of the rationale behind recent US measures such as PADFA and the Department of Justice's bulk sensitive data rule, and it gives public-safety and national-security authorities a direct stake in how these logs are retained and protected.¹⁸⁷

Two very different kinds of failure have to be kept apart, because they are guarded against in different ways. The **first**, and by far the more common, is **compromise of the user's own account**. A phished password, a reused credential, malware on the user's device, or a stolen session lets an attacker sign in as the user. To the provider's systems this looks like a legitimate login, so the defences built to keep outsiders out, encryption, database access controls, network segmentation, do nothing against it: the intruder is already inside, holding the user's key. Account-security features such as multi-factor authentication and anomaly detection reduce how often this happens, but they do not change what is exposed once it does. This is why account takeover is the central case for chatbot confidentiality rather than a footnote to it (see Section 4.2.5): it is the likeliest route to exposure, and it is precisely the route that hardening the provider's infrastructure cannot close. The only protection that bites is not holding the data in the first place.

The **second** kind of failure is **a breach of the provider's own systems**: a database left open to the internet, a bug that serves one user's content to another, a compromised vendor in the processing chain, or a malicious insider (see Sections 4.2.1 to 4.2.4 and 4.4). Here the provider's technical safeguards matter, and matter greatly. Encryption at rest and in transit, strict access controls, segmentation, key management, vendor contracts under GDPR Article 28, breach notification, and incident response together determine whether an intrusion reaches plaintext content at all, and how far it travels if it does. A breach of a short-lived, encrypted, siloed, minimally accessible dataset has a very different blast radius from a breach of years of searchable conversations, memory profiles, uploaded files, and agentic action logs.

The two failure modes converge on one conclusion. Against a breach of the provider, retention and architecture together set the damage; against account takeover, retention sets it alone. The cybersecurity lesson and the data-minimisation lesson developed throughout this study are therefore the same lesson: the safest chatbot log is the one that does not exist, does not persist, or cannot be meaningfully connected back to a person.¹⁸⁸

¹⁸⁶ See GDPR [Article 4\(12\)](#) for the definition of a personal data breach; [Articles 33](#) and [34](#) govern notification to supervisory authorities and communication to affected data subjects.

¹⁸⁷ On the national-security dimension of personal-data exposure, see the [Protecting Americans' Data from Foreign Adversaries Act of 2024](#) (PADFA), enacted as Division I of Pub. L. No. 118-50, and the Department of Justice's [Data Security Program](#) implementing Executive Order 14117 of 28 February 2024, "[Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern](#)", final rule of 8 January 2025, codified at 28 C.F.R. pt. 202. The Order was driven in part by the concern that adversary states could use AI to mine bulk sensitive data for intelligence value, a risk that conversational logs, dense with a user's intentions, affiliations and daily routines, only sharpen.

¹⁸⁸ Meta's *Incognito Chat with Meta AI*, announced on 13 May 2026, illustrates this principle in deployed form. The TEE-based Private Processing architecture decrypts user content only inside the enclave and only for the duration of the request; the plaintext does not leave the enclave, is not written to persistent storage, and is purged from memory when the session ends. The breach surface for the corresponding category of conversations is therefore much narrower than for the conventional cloud-AI architectures whose breach record this chapter reviews: there is no years-long, searchable, indexed corpus on the provider's servers to be exfiltrated, abused by an insider, or compelled by external process. The trade-offs (the loss of provider-side moderation visibility, the loss of memory-augmented user experience, the residual hardware-trust assumptions documented by external auditors) are analysed in Christakis and Swire (cited above n 11).

4.2. The public record: warning incidents, not yet a full catastrophe

The public record already contains several incidents that illuminate the risk. They should be treated carefully. Some are true security breaches. Others are design failures, accidental disclosures, credential-compromise incidents, or examples of enterprise misuse. They do not prove that any major chatbot provider has suffered a catastrophic exposure of full conversation histories. But they do show that the risk is no longer theoretical. The incidents can be grouped into five categories that correspond to the failure modes identified at the outset of this chapter.

4.2.1. Infrastructure vulnerabilities and software bugs

The first major ChatGPT privacy incident occurred in March 2023. OpenAI took ChatGPT offline after a bug in the open-source Redis client library allowed some users to see titles from another active user's chat history. OpenAI later reported that, in limited circumstances, the first message of a newly created conversation could also have been visible in another user's history. The same bug may also have exposed payment-related information of 1.2% of ChatGPT Plus subscribers active during a specific nine-hour window, including name, email address, payment address, credit card type, last four digits, and card expiration date. The episode was not a mass leak of full conversation histories, but it remains important because it showed that even a sophisticated provider could suffer a cross-user data exposure caused by ordinary software infrastructure.¹⁸⁹

The most direct chatbot-log exposure identified in the public record is the DeepSeek database incident of January 2025. Wiz researchers reported that they found a publicly accessible ClickHouse database belonging to DeepSeek that required no authentication and allowed full control over database operations. According to Wiz, the exposed data included over one million lines of log streams containing chat history, secret keys, backend details, and other sensitive information. Reuters reported that the logs captured user instructions to DeepSeek's free AI assistant and that DeepSeek secured the data after being alerted by Wiz. Whether anyone else accessed the database remains unknown.¹⁹⁰ For this study, the DeepSeek incident is significant for two reasons. First, it is not merely a hypothetical concern about future chatbot-log breaches: chat histories were reportedly exposed. Second, it shows that the relevant data is not only user content. Operational logs may contain prompts, backend details, API keys, and infrastructure information, creating both privacy risk and security escalation risk.

The risk is not limited to the five providers examined in this study. The Chat & Ask AI incident of January 2026 demonstrated that third-party "wrapper" apps, which route conversations through OpenAI, Anthropic, and Google models, can create their own exposure surfaces. Chat & Ask AI, a popular mobile app with more than 50 million users, was found by an independent security researcher to have a misconfigured Google Firebase backend that reportedly exposed approximately 300 million messages from more than 25 million users. The exposed messages included deeply personal and disturbing content: users asking how to harm themselves, how to write suicide notes, and how to manufacture controlled substances. The developer, Turkish company Codeway, secured the database within hours of being notified, but the incident illustrates a critical point: the confidentiality of a chatbot conversation depends not only on the policies of the underlying model provider but also on the security practices of every intermediary through which the conversation passes.¹⁹¹ A user interacting with Claude through a third-party app may assume they are protected by Anthropic's security infrastructure, when in fact their conversation is stored, processed, and potentially exposed by an entirely separate entity with different security standards.

¹⁸⁹ OpenAI, "[March 20 ChatGPT outage: Here's what happened](#)," 24 March 2023.

¹⁹⁰ Wiz Research, "[Wiz Research Uncovers Exposed DeepSeek Database Leak](#)," 29 January 2025; Reuters, "[Sensitive DeepSeek data exposed to web, Israeli cyber firm says](#)" 29 January 2025; The Verge, "[DeepSeek database left user data, chat histories exposed for anyone to see](#)" 30 January 2025.

¹⁹¹ 404 Media, "Massive AI Chat App Leaked Millions of Users' Private Conversations," 29 January 2026; CyberGuy / Fox News, "[Millions of AI chat messages exposed in app data leak](#)," 5 February 2026; Malwarebytes, "[AI chat app leak exposes 300 million messages tied to 25 million users](#)," 9 February 2026.

4.2.2. Sharing design failures and accidental publication

A different category of exposure arose in 2025 through public-sharing interfaces. OpenAI briefly allowed users to make shared ChatGPT conversations discoverable by search engines. After reports that sensitive shared conversations were appearing in Google search results, OpenAI removed the feature. OpenAI's Chief Information Security Officer described it as a short-lived experiment that created too many opportunities for users to accidentally share things they did not intend to share, and said the company was working to remove indexed content from relevant search engines.¹⁹² This was not a provider-side breach in the narrow sense: users had to use the sharing flow and opt into discoverability. But the incident belongs in this chapter because it shows how easily the boundary between “private chat”, “shared link”, and “public web page” can collapse when interface friction is insufficient. A conversation that feels private can become searchable, cached, and copied.

The Grok episode that followed a few weeks later showed the same design risk at much greater scale. In August 2025, Forbes reported that more than 370,000 conversations with xAI's Grok had been indexed by search engines, including Google, Bing and DuckDuckGo, after users created shareable links to their chats. Reporting indicated that the indexed material included medical and psychological questions, business details, passwords or credentials, and prompts or outputs concerning illegal or harmful activities.¹⁹³ Unlike OpenAI's analogous experiment, which reportedly resulted in roughly 4,500 indexed conversations before the feature was withdrawn, the Grok incident operated at the scale of hundreds of thousands of chats. The episode illustrates that human access is not the only vector through which private conversations can escape the provider's perimeter: design failures in sharing, collaboration, or export features can produce exposure at scale, particularly where privacy-by-design principles are not embedded into feature development.

The Meta AI public-feed controversy illustrates the same point in a more social design. Meta's standalone AI app included a Discover feed displaying user interactions that users could share publicly. Press reports found that some users appeared to be publishing highly personal AI interactions, including medical, intimate, religious, emotional, and legal material, apparently without fully understanding the public visibility of what they were sharing. Business Insider later reported that Meta added a warning before posting to the public feed, after reporting showed apparently accidental publication of personal information.¹⁹⁴ Again, this is not a conventional intrusion by an external attacker. It is closer to privacy by confusing design. But from the affected user's perspective, the difference may matter little: the result is exposure of sensitive chatbot content to strangers.

4.2.3. Client-side and supply-chain exfiltration

Perhaps the most disturbing exposure vector documented in the public record to date involves not a provider breach or a design flaw, but a deliberate data-harvesting operation conducted through browser extensions marketed as privacy tools. In December 2025, security researchers at Koi Security discovered that a suite of browser extensions, most prominently Urban VPN Proxy (a Chrome extension with over 6 million users and Google's “Featured” badge), had been intercepting prompts, responses, conversation identifiers and metadata from major AI platforms, including ChatGPT, Claude, Gemini, Copilot, Perplexity, DeepSeek, Grok and Meta AI. The harvesting code was introduced through a silent auto-update in July 2025, meaning users who had installed the extension for its VPN functionality were upgraded to a surveillance version without any new consent request. The data was transmitted to Urban VPN infrastructure and, according to Koi's analysis of the corporate structure and privacy

¹⁹² Fortune, [“OpenAI will no longer let ChatGPT conversations be discoverable on Google,”](#) 5 August 2025; Business Insider, [“OpenAI quickly rolled back a new feature that allowed users to make private conversations with ChatGPT searchable,”](#) 31 July 2025.

¹⁹³ Forbes, [“Elon Musk's xAI Published Hundreds Of Thousands Of Grok Chatbot Conversations,”](#) 20 August 2025, reporting that more than 370,000 Grok conversations were indexed by search engines; Fortune, [“Thousands of private user conversations with Elon Musk's Grok AI chatbot are now publicly available on Google,”](#) 22 August 2025; Malwarebytes, [“Grok chats show up in Google searches,”](#) 22 August 2025.

¹⁹⁴ Business Insider, [“We uncovered how Meta's AI app was full of accidental public posts that were really personal,”](#) June 2025; Washington Post, “Meta AI users confide on sex, God and Trump. Some don't know it's public,” 13 June 2025; Malwarebytes, “Your Meta AI chats might be public, and it's not a bug,” 13 June 2025.

documentation, shared within an ecosystem connected to BiScience, a data broker. Across eight related extensions, more than 8 million users were affected.¹⁹⁵

The Koi Security findings were subsequently confirmed and extended by independent researcher Lee Dryburgh, who in early 2026 documented that chatbot conversation data harvested through such browser extensions was being stored in searchable vector databases and sold to commercial customers. Dryburgh reported finding conversations containing real names, dates of birth, medical record numbers, diagnosis codes, and clinical notes, including instances where healthcare workers had pasted real patient data into AI chatbots. The data was pseudonymised with SHA-256 hashes,¹⁹⁶ but the verbatim content of conversations was fully searchable, and many prompts contained enough contextual detail to be re-identifiable.¹⁹⁷ The browser-extension vector is analytically important because it bypasses all provider-side security measures. It does not matter how well OpenAI or Anthropic encrypt their databases, manage access controls, or vet their vendors if the conversation is intercepted at the browser level before it reaches the provider's servers. The risk is invisible to the user and to the provider. It represents a supply-chain attack on the trust relationship between user and chatbot, and operationally a man-in-the-browser one: the conversation is read inside the browser, in cleartext, before transport encryption or any provider safeguard can apply.

4.2.4. Tracker and telemetry leakage from providers' own platforms

A research disclosure published on 4 May 2026 by researchers at IMDEA Networks documented a further exposure vector that is distinct from all of the above: leakage of conversation URLs, chat titles or topics, user identifiers, and related metadata to third-party advertising, analytics and support services embedded in the providers' own web interfaces. The LeakyLM study tested four major platforms, Perplexity, Anthropic's Claude, OpenAI's ChatGPT, and xAI's Grok, and found that all four transmitted some combination of conversation URLs, page titles, user identifiers, cookies, email addresses or email hashes, and usage metadata to external services, including Meta, Google, TikTok, Datadog, Intercom and others. In some configurations, the researchers found that these transmissions occurred regardless of cookie choices; in others, they were triggered only after acceptance of non-essential cookies.¹⁹⁸

The privacy implications depend on the interaction between two factors: what data is transmitted and whether the conversation URL gives access to the conversation content. On Grok, the researchers found that conversation permalinks are publicly accessible by default, meaning that any tracker receiving the URL can in principle access and read the full conversation. Meta Pixel, Google Analytics, DoubleClick, and TikTok all received Grok conversation URLs. TikTok's tracker additionally received verbatim message content exposed through Open Graph metadata in conversation screenshot images. For Claude, the researchers found that, when non-essential cookies were accepted, Anthropic's infrastructure forwarded certain user events server-side to eleven advertising and analytics platforms, including Meta, LinkedIn, TikTok and Google; they also found that Intercom transmitted conversation URLs on authenticated sessions regardless of cookie choices. For ChatGPT, Google Analytics received conversation URLs and page titles (which reflect the chat topic and can reveal highly sensitive

¹⁹⁵ Koi Security, "[8 Million Users' AI Conversations Sold for Profit by 'Privacy' Extensions](#)," December 2025; Ars Technica, "[Browser extensions with 8 million users collect extended AI conversations](#)" 17 December 2025; Malwarebytes, "[Chrome extension slurps up AI chats after users installed it for privacy](#)," 18 December 2025.

¹⁹⁶ SHA-256 is a standard cryptographic hash function, specified by NIST in FIPS 180-4 and documented for Internet use in IETF RFC 6234. Hashing the identifiers attached to a conversation is a recognised pseudonymisation measure, but it protects only those identifiers. Where the conversational content itself is retained in cleartext and remains searchable, as it was here, a robust hash on the surrounding identifiers adds little, because the text speaks for itself.

¹⁹⁷ Lee Dryburgh, as reported in The Register, "[Chatbot data harvesting yields sensitive personal info](#)," 3 March 2026, documenting that harvested chatbot conversation data included "healthcare workers pasting real patient data into AI chatbots" with conversations containing "real names, dates of birth, medical record numbers, and diagnosis codes" stored in searchable commercial databases. See also Koi Security, "8 Million Users' AI Conversations Sold for Profit by 'Privacy' Extensions," December 2025, *op. cit.*. The use of SHA-256 hashes is industry-standard.

¹⁹⁸ LeakyLM, "[Your AI Assistant Is Leaking Your Conversations](#)," IMDEA Networks research disclosure, 4 May 2026. The research team tested Perplexity, Anthropic's Claude, OpenAI's ChatGPT, and xAI's Grok across combinations of authentication state, cookie consent, account tier, and privacy mode. Findings were submitted to relevant data protection authorities on 13 April 2026. xAI was notified on 17 April 2026; no response had been received at the time of publication. Meta AI, Microsoft Copilot, and Google Gemini were excluded from the LeakyLM analysis because they act both as LLM providers and as third-party trackers, falling into a different threat model.

information: theoretical examples given by Jorge Garcia Herrero include “Tips for depression”, “Aggressive tax deductions”, “My partner cheated on me¹⁹⁹”) for free logged-in users regardless of cookie consent.

The server-side tracking dimension deserves particular attention. Traditional privacy controls available to users (rejecting cookies, installing ad blockers, using VPNs) operate at the browser level. Server-side tracking bypasses all of them: the data is transmitted from the provider’s own infrastructure to third-party advertising platforms through server-to-server connections that the user’s browser never sees and cannot intercept. As Jorge Garcia Herrero observed in his legal analysis of the LeakyLM findings, even users who take the trouble to reject cookies and install privacy extensions cannot prevent these data flows.²⁰⁰

The LeakyLM findings are significant for this study²⁰¹ in three respects.

First, they show that the third-party data flows documented in Part 1 of this study extend beyond operational sharing with processors and vendors (which may be lawful and necessary) to include transmission of conversation identifiers to advertising, analytics, observability, and support services, some of which operate within broader behavioural-profiling ecosystems.

Second, and more concretely for the breach analysis in this chapter, the combination of weak permalink access controls and tracker-mediated URL dissemination creates a vulnerability: a conversation that the user believes is private may be accessible to any entity that receives the permalink.

Third, from a GDPR perspective, as Garcia Herrero argues, the mere act of making conversation data available to third-party trackers constitutes a processing operation that is distinct from the service the user contracted for, and that requires its own legal basis, its own transparency, and its own justification under the CJEU’s *Meta Platforms Ireland* standard (C-757/22, paragraph 54), which requires the controller to inform the data subject specifically of the purposes and recipients of each processing operation. On that view, generic references in privacy policies to “user content”, “conversations”, “service interaction information” or “business partners” may be insufficient if they do not allow users to understand that conversation URLs, titles, identifiers, or related metadata are transmitted to specific third-party advertising, analytics or support services. Whether regulators will adopt that analysis remains to be seen, but the LeakyLM findings make the transparency issue concrete. A class action lawsuit filed in March 2026 in the US (*Doe v. Perplexity AI, Meta Platforms, Google*, Case 3:26-cv-02803) alleges that Perplexity’s privacy controls, including its incognito mode, failed to prevent the

¹⁹⁹ See the following footnote.

²⁰⁰ J. Garcia Herrero, “[Every prompt you make, AI’ll be watching you](#),” Zero Party Data, 5 May 2026, providing a legal analysis of the LeakyLM findings under the GDPR. Garcia Herrero argues that the four providers’ transparency obligations under Article 13 GDPR and the CJEU’s *Meta Platforms Ireland* standard (Case C-757/22, paragraph 54) require specific information about the purposes and recipients of each processing operation, and that the generic language used in the providers’ privacy policies does not meet this standard. On server-side tracking as a mechanism that bypasses all user-side privacy controls, see also C. Santos et al., “[The Devil is in the Details: Detection, Measurement and Lawfulness of Server-Side Tracking on the Web](#),” Proceedings on Privacy Enhancing Technologies (PoPETs), 2024.

²⁰¹ A related exposure at the same live-interaction layer has so far attracted even less scrutiny. The transmissions documented by LeakyLM concern, principally, identifiers, conversation URLs, chat titles that may reveal conversation topics, tracking cookies and other data sent to embedded analytics, advertising or support services. A distinct question arises once a chatbot operates agentially or through tool use. When the system composes a call to an external tool, a web search service, a third-party database or a connected application, the user may no longer be the person who writes or even sees the parameters of the disclosure. Those parameters may be generated from the conversational context, including sensitive information the user supplied to the assistant in confidence. Content disclosed to the assistant can therefore be repackaged into a query or API call sent to a separate tool provider, without a discrete act of disclosure by the user and without the user necessarily appreciating that the relevant information has left the chatbot environment. The analysis advanced above applies with particular force in that setting. A disclosure to a separate recipient is itself a processing operation, and is subject to the GDPR requirements of lawfulness, fairness, transparency, purpose limitation and data minimisation. At a minimum, users should be able to understand which categories of tools or recipients may receive their data, for what purposes, on what legal basis, and with what safeguards. This is also the logic of C-757/22: under Article 13(1)(c) and (e) GDPR, the controller must inform the data subject of the purposes of the processing, the legal basis for it, and the recipients or categories of recipients of the data, with Article 12(1) governing the form in which that information is provided ([Case C-757/22](#), *Meta Platforms Ireland Ltd v Bundesverband* judgment of 11 July 2024, para. 54). The problem is even sharper where the payload is not merely a cookie, URL or identifier, but substantive conversational content or a model-generated distillation of it. What is actually transmitted in these agentic calls, across providers, tool ecosystems and connector types, has not, to my knowledge, been the subject of systematic empirical study comparable to recent work on web tracking in chatbot interfaces. It should be.

transmission of conversation data to Meta and Google.²⁰² Perplexity reportedly removed the Meta Pixel on 3 April 2026, shortly after the lawsuit was filed.

The issue has since moved onto the European regulatory agenda. On 27 May 2026 the Spanish Data Protection Agency (AEPD) sent the European Data Protection Board an informative note on the LeakyLM study, asked the Board to circulate it to all European supervisory authorities for their knowledge and assessment, and pressed for the matter to be put on the agenda of the EDPB plenary of 8 and 9 June 2026.²⁰³ The AEPD stressed the need to address the study's results in coordination with its European counterparts, by reference to the GDPR's one-stop-shop mechanism, and relayed the study's identification of three main problems: the exposure of conversation permalinks to third-party trackers, the linkability of those interactions to user identities through tracking mechanisms, and privacy controls and policies that may not accurately reflect the actual data flows. Two features of the move matter here. It turns collective European regulatory attention to the live-interaction layer, the data flows that occur during the conversation itself, which is the pathway this chapter examines. And because the AEPD presents the practices as industry-wide and seeks coordinated treatment rather than action against any single provider, any response at EU level would have cross-provider effect.

4.2.5. Credential compromise and enterprise leakage

This final category covers two failures that meet at the endpoint rather than in the provider's systems, and that are worth keeping apart. One is credential and device compromise, in which an outside attacker reaches the conversation through the user's own machine. The other is enterprise leakage, in which authorised users hand confidential material to the chatbot themselves. The first is an attack, the second a governance failure, and neither is stopped by anything the provider does to harden its own infrastructure, which is precisely why they close this survey.

In June 2023, Group-IB reported that it had identified 101,134 stealer-infected devices with saved ChatGPT credentials in information-stealer logs traded on illicit marketplaces between June 2022 and May 2023. The issue was not a breach of OpenAI's infrastructure. The credentials were harvested from users' infected devices. But the effect of a compromised chatbot account can be unusually intrusive. If the account preserves conversation history, memory, uploaded files, or connected services, the attacker may gain access not only to a login credential but to a record of the user's private reasoning over time.²⁰⁴ In the age of memory and agentic assistants, account takeover becomes a pathway into a personal dossier.

Information-stealer malware of the kind Group-IB tracked is opportunistic and works at scale. At the targeted end of the same spectrum sits commercial spyware. Tools such as NSO Group's Pegasus can compromise a smartphone through zero-click exploits that require no action by the user at all, and can read whatever the device holds or displays. A chatbot conversation on such a device is exposed along with everything else, whatever the account security, the transport encryption, or the quality of the provider's server-side controls. The documented targeting of journalists, lawyers, activists, and officials makes this a live risk for exactly the high-stakes users whose chatbot logs are most revealing.²⁰⁵

Finally, the Samsung incident remains the canonical example of enterprise "shadow AI" leakage. In 2023, Samsung employees reportedly entered confidential source code, internal testing code, and meeting material into ChatGPT while using it for work tasks. The incident led Samsung to restrict or ban employee use of generative AI tools for a period and accelerated enterprise concern about

²⁰² See Ars Technica, "[Perplexity's incognito mode is a sham, lawsuit says](#)," April 2026.

²⁰³ Agencia Española de Protección de Datos, "[La Agencia promueve ante las autoridades europeas de protección de datos que se estudie si algunos sistemas de IA permiten a terceros acceder a las conversaciones](#)" [The Agency calls on the European data protection authorities to study whether some AI systems allow third parties to access conversations], press release, 27 May 2026.

²⁰⁴ Group-IB, "[Group-IB discovers 100K+ compromised ChatGPT accounts](#)," 20 June 2023; The Hacker News, "[Over 100,000 Stolen ChatGPT Account Credentials Sold on Dark Web Marketplaces](#)" 20 June 2023.

²⁰⁵ On the capabilities and documented use of mercenary spyware, see Amnesty International, [Forensic Methodology Report: How to Catch NSO Group's Pegasus](#) (2021), released alongside the Pegasus Project investigation, and the forensic reporting of the [Citizen Lab](#). Pegasus has been delivered through zero-click exploits against the devices of journalists, lawyers, human-rights defenders, and political figures in numerous countries.

uncontrolled use of public AI systems.²⁰⁶ This was not a breach of a chatbot provider. It was a breach of organisational confidentiality through intended use of a chatbot. Its lesson is nevertheless central: the biggest leak channel may not be hackers breaking into the chatbot. It may be authorised users pasting secrets into it because the tool is useful.

4.3. The new risk introduced by agents

The cases above mostly concern stored conversations, sharing interfaces, credentials, or logs. Agentic AI adds a further layer. A chatbot that merely answers a prompt is a confidentiality risk because it receives and stores sensitive text. An AI agent that reads email, browses websites, manages files, interacts with calendars, books travel, drafts messages, or invokes plugins is a security risk of a different order. It can act.

The risk is that **agents collapse boundaries that traditional security models kept separate**. They read untrusted external content, combine it with trusted internal context, and may act through the user's permissions. That architecture creates the conditions for indirect prompt injection: malicious instructions hidden in a webpage, email, document, calendar invitation, or other external source may be interpreted by the agent as instructions to follow, rather than as content to summarise.

Microsoft's own security materials describe indirect prompt injection as a risk that arises when an LLM processes untrusted data and misinterprets crafted content as instructions.²⁰⁷ The EchoLeak vulnerability in Microsoft 365 Copilot, reported in 2025 as CVE-2025-32711, showed the stakes. Researchers described a zero-click prompt-injection vulnerability that could enable remote unauthenticated data exfiltration through a crafted email in a production AI assistant environment. Microsoft remediated the vulnerability server-side before public disclosure, and no exploitation in the wild was identified. The details are technical, but the governance lesson is simple: an AI assistant that has access to internal documents, emails, and enterprise context can become a bridge between the attacker's input and the user's data.²⁰⁸

Consumer chatbot providers are moving toward the same terrain. Memory, uploaded files, connectors, browsing, plugins, custom GPTs, workplace integrations, health agents, and personal agents all increase usefulness by giving the system more context and more authority. But they also increase the consequences of compromise. **If an attacker can manipulate an agent**, steal an access token, compromise a connected account, or induce the agent to disclose context, the exposed material is no longer limited to the current prompt. It may include email, documents, calendar entries, uploaded files, health records, browsing sessions, contact lists, or memory summaries.

The OWASP Top 10 for LLM Applications captures this shift in security vocabulary. The 2025 framework identifies prompt injection, sensitive information disclosure, supply-chain vulnerabilities, vector and embedding weaknesses, and excessive agency as major categories of LLM risk.²⁰⁹ These are the technical counterpart of the legal and privacy problems analysed in this study. "Sensitive information disclosure" maps onto the exposure of chatbot logs. "Excessive agency" maps onto an assistant acting beyond what the user or developer expected. "Vector and embedding weaknesses" map onto the risk that memory stores, retrieval systems, and uploaded knowledge bases become new leakage surfaces. The more the chatbot becomes an operating layer for personal and professional life, the less adequate it is to think of security as merely protecting a chat transcript.

²⁰⁶ CIO Dive, "[Samsung employees leaked corporate data in ChatGPT](#)," 7 April 2023; Forbes, "[Samsung Bans ChatGPT Among Employees After Sensitive Code Leak](#)" 2 May 2023.

²⁰⁷ Microsoft Security, "[Detecting and analyzing prompt abuse in AI tools](#)," 12 March 2026; Microsoft Security Response Center, "[How Microsoft defends against indirect prompt injection attacks](#)" 29 July 2025.

²⁰⁸ Pavan Reddy and Aditya Sanjay Gujral, "[EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System](#)," arXiv, 2025; The Hacker News, "[Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction](#)" 12 June 2025.

²⁰⁹ OWASP, [Top 10 for LLM Applications and Generative AI](#), 2025.

4.4. Insider access and vendor exposure

External attackers are only one part of the problem. Part 1 of this study emphasised that human review is structural across chatbot services and that Sealed Mode should include minimised routine human access, strict escalation rules, least-privilege controls, and immutable audit logs.²¹⁰ That recommendation is not only about privacy. It is also about breach prevention.

Insider risk has several forms. The simplest is malicious access by an employee or contractor. More common is excessive legitimate access: too many teams, support personnel, trust-and-safety reviewers, annotators, engineers, vendors, or auditors can see too much data for too long. A third form is workflow exposure: sensitive conversations may enter ticketing systems, debugging environments, analytics dashboards, quality-review datasets, or vendor platforms that were not designed for the sensitivity of the underlying content. The provider may be secure at the core while the periphery remains fragile.

The vendor dimension is particularly important because modern AI services are not operated by a single entity in isolation. They rely on cloud infrastructure, customer-support systems, security tools, annotation pipelines, abuse-detection contractors, legal vendors, and analytics systems. Part 1 insisted that operational sharing with processors and vendors is not inherently improper and is legally different from selling data. Under GDPR Article 28, processors are bound by documented instructions, confidentiality, and appropriate security measures. But it also noted that users often cannot assess who may access their conversations.²¹¹ Chapter 3 added the litigation analogue: once chatbot logs are produced in discovery, they may reside with law firms, experts, e-discovery platforms, cloud repositories, review databases, and managed-service vendors, each of which can become a target or failure point. That analysis applies equally to ordinary operations. The confidentiality of chatbot logs is only as strong as the weakest system that stores, indexes, reviews, exports, backs up, or analyses them.

This is why a breach chapter must be connected to retention and segmentation. If sensitive chats are copied into multiple operational systems, **the provider must secure not one database but an ecosystem**. If prompts are embedded into logs, analytics tables, model-evaluation datasets, or safety-review queues, deletion from the chat interface may not delete the operational residue. If vendors receive plaintext where derived or redacted data would suffice, the attack surface expands unnecessarily. The core security question is therefore: “where did the conversation travel, who could read it, what copies were made, and how long did they persist?”

The question of breach risk is therefore inseparable from the question of deletion fidelity: a user-facing deletion control is not enough if copies, embeddings, logs, analytics events, support tickets, or safety-review records continue to preserve the substance or identifiers of the conversation in other systems.

4.5. The legal lens: breach notification is necessary but not enough

European data protection law already provides an important baseline.²¹² Article 32 GDPR requires controllers and processors to implement security appropriate to risk, taking account of the state of the art, implementation costs, processing context, and the likelihood and severity of risks to individuals. Articles 33 and 34 create notification duties to supervisory authorities and, in high-risk cases, to affected individuals.²¹³ These rules are directly relevant to chatbot logs. A breach of conversations involving health, sexuality, children, legal matters, immigration status, political opinion, credentials, or trade secrets will often carry risks that go far beyond ordinary account data. These notification duties rest on a familiar economic rationale. The gains from collecting, retaining, and connecting personal data accrue largely to the controller, while much of the loss from a breach falls on the individuals concerned and on third parties. Economic analysis has long shown that, under these conditions, a market left to itself tends

²¹⁰ See T. Christakis, [You Trust Your Chatbot With Everything. Should You? Part 1](#), *op. cit.*.

²¹¹ Part 1’s discussion of operational sharing emphasises that processor relationships can be necessary and legally constrained, but that users deserve clearer recipient transparency, including which categories of vendors and teams can access plaintext chat content. See Part 1, *op. cit.*.

²¹² Other jurisdictions have similar cybersecurity laws. For instance, [at least 25 of the 50 states](#) in the U.S. address data security practices of private sector entities.

²¹³ GDPR Articles [32](#), [33](#) and [34](#) address security of processing, notification of personal data breaches to supervisory authorities, and communication of high-risk breaches to affected data subjects.

to underprovide security relative to what would be efficient for society, which is one reason the law intervenes to require notification in the first place.²¹⁴

But breach notification comes after failure. For chatbot systems, the harder question is architectural, and it falls to management as much as to law: what level of retention, internal access, vendor distribution, logging, and agentic permission is appropriate before an incident occurs? A provider cannot credibly claim that a breach of chatbot logs is low risk simply because the exposed dataset lacks names or account IDs. As Chapter 3 explained in the discovery context, de-identification leaves substantive secrecy exposed where the content itself contains unique contextual details. The same is true in a breach. A “de-identified” conversation about a rare diagnosis, a pending acquisition, a child’s school, or a legal dispute may still be re-identifiable to those with context.

The regulatory response should therefore move beyond ordinary breach-notification compliance. Chatbot providers should treat conversation logs, memory files, uploaded documents, and agentic action logs as high-sensitivity datasets by default. That does not mean that every conversation contains special-category data. But it means that the system is designed to receive such data at scale, and that the provider cannot know in advance which exchanges will be highly sensitive. The appropriate baseline is **risk-calibrated hardening**, not reactive classification after exposure.

4.6. Conclusions and tentative recommendations

The public record does not yet show a single catastrophic breach in which a major provider’s entire consumer chatbot corpus was exfiltrated. That is good news. But it should not be comforting. The incidents already documented are enough to establish the risk pattern. A software dependency exposed ChatGPT titles and limited payment data. A DeepSeek database reportedly exposed chat history, API keys, and operational logs. A third-party wrapper app exposed 300 million messages from 25 million users through a misconfigured Firebase backend. Shared-link and public-feed designs turned private-seeming conversations into searchable or publicly visible artefacts at the scale of hundreds of thousands. Browser extensions marketed as privacy tools silently intercepted and sold complete AI conversations from 8 million users to a data broker. Researchers reported that trackers embedded in the providers’ own interfaces transmitted conversation URLs and user identifiers to advertising networks including Meta, Google, and TikTok, in some configurations regardless of cookie consent. Stolen credentials created pathways into preserved chat histories. Enterprise employees leaked secrets by pasting them into public AI tools. AI-agent vulnerabilities demonstrated that prompt injection can become data exfiltration when the system has access to external context and delegated permissions.

The lesson is structural. Chatbot confidentiality can fail even when no court orders disclosure, no agency compels access, and no provider voluntarily refers a conversation. It can fail because data is retained longer, copied more widely, logged more casually, shared more opaquely, indexed more easily, intercepted more silently, and connected more deeply to other systems than a risk-based assessment would justify. The same features that make chatbots valuable (memory, continuity, file access, connected apps, human review, safety monitoring, and agentic action) also determine the blast radius of a breach.

Sealed Mode is therefore not only a privacy feature. It is a **resilience architecture**. By combining strict retention, no training, no advertising, siloed personalisation, minimised human access, strong auditability, and cryptographic or infrastructural hardening, it reduces both ordinary misuse and extraordinary exposure. It will not eliminate all risk. But it would make the most sensitive conversations less available, less searchable, less replicated, and less vulnerable to the **predictable failures of complex systems**.

²¹⁴ On why markets tend to underprovide data security relative to the social optimum, given that much of the cost of a breach falls on parties other than the controller, see Paul M. Schwartz and Edward J. Janger, “[Notification of Data Security Breaches](#)”, 105 Mich. L. Rev. 913, 928 (2007). These externalities are a principal economic justification for breach-notification statutes.

Tentative Recommendations on Cyberthreats, Insider Risks and Data Breaches

#	Recommendation
15	Treat chatbot logs as high-sensitivity data by default. Providers should classify consumer chatbot logs, memory files, uploaded documents, and agentic action logs as high-sensitivity datasets for security purposes, even when no special-category data has yet been identified in any particular conversation. The reason is that the system predictably receives sensitive data of every kind and concentrates it through persistent memory profiles and agentic action logs that have no equivalent in older services. Where input of that kind is predictable, the security baseline should match it rather than the lowest-sensitivity conversation the system might contain, in the same way that data-protection law attaches heightened requirements to categories of data it defines as sensitive.
16	Reduce the breach blast radius through minimisation, segmentation, and Sealed Mode as a security control. The safest chatbot thread is the one that does not exist or does not persist. Providers should minimise default retention, separate chat history from memory profiles, isolate uploaded files from ordinary conversation logs, and avoid unnecessary replication into analytics, support, testing, or model-evaluation systems, including any flow of conversation metadata to third-party services not strictly necessary for product operation. The Sealed Mode proposed in Part 1 should be designed not only as a privacy feature but as a security architecture, with strict retention, sharply limited human access, just-in-time authorisation, immutable audit logs, and stronger cryptographic or infrastructural protections.
17	Harden access controls across insiders, vendors, and AI agents. The same least-privilege logic should apply across all three categories. For insiders and vendors, providers should adopt just-in-time approval, segregation of duties, logging of access to sensitive conversation datasets, and contractual controls for any third party that processes chatbot content. For AI agents acting on the user's behalf, providers should ensure short-lived tokens, explicit user confirmation for high-risk actions, and strong controls against indirect prompt injection. Agents that can read email, files, health records, or enterprise repositories should be treated as security-sensitive applications.
18	Design sharing interfaces and breach response with chatbot-data sensitivity in mind. The ChatGPT indexing episode and the Meta AI Discover feed controversy showed that sharing interfaces designed for ordinary content can produce serious exposure when applied to confidential conversations. Sharing features, public-link options, and collaboration modes should be designed on the assumption that users will misunderstand visibility settings: the default should be private, public sharing should require clear friction and easy revocation, and indexing should not occur by default. Where chatbot logs or memory profiles are nevertheless exposed, breach assessment should look at the substance and context of the conversations, not only at identifiers, and notifications should explain what categories of content may have been affected and what remedial steps users can take. For chatbot data, a technically narrow breach may still be personally devastating.
19	Publish AI-specific security and breach transparency. Transparency reports should not be limited to government requests and CSAM reports. Providers should publish aggregate information about material security incidents affecting chatbot data, to the extent compatible with security and legal obligations. Users and regulators need to understand not only how often the state asks for chatbot logs, but also how often the systems holding those logs fail.

5. Conclusions and Consolidated Recommendations

The four chapters of this study have examined four pathways by which a chatbot conversation can leave the confidential space the user expects: proactive corporate disclosure, compelled government access, discovery, and security failure. Each pathway operates under a different legal framework, raises different governance questions, and produces different forms of harm. The aim of these conclusions is to draw out what the four pathways have in common, to be explicit about what is familiar in form and what is new about the chatbot context, to register the architectural development that occurred during the final stage of drafting, and to consolidate the tentative recommendations of the four chapters in a single condensed form.

5.1. The four pathways and what they share

Each of the four pathways examined in this study is governed by its own legal framework and has its own analytical character. Three features nonetheless run across all of them.

The first is a **common dependency**. All four pathways become possible, or become more damaging, because chatbot conversations are retained, linked to accounts, reviewed, logged, indexed, personalised, connected to other services, or stored in forms that can later be searched or disclosed. These are not only legal and contractual choices but management and operational ones, made by product, engineering, and senior leadership, about what data exists, for how long, and in what form. The same choices that make a chatbot useful, namely memory, continuity, file access, connected apps, human review, safety monitoring, and agentic action, are the ones that set the surface across which all four pathways operate. Privacy features in Part 1 turn out, in Part 2, to be the determinants of how seriously each of the four pathways exposes the user.

The second is **convergence rather than substitution**. The four pathways do not compete; they coexist. A user whose conversation is the subject of a discovery order may also be the subject of a proactive emergency referral if it contains material that meets the relevant threshold. The same retention practices that support post-incident law-enforcement cooperation also expand the dataset available to opposing counsel in discovery and to attackers in a breach. The same memory feature that improves the product creates structured longitudinal profiles that are surveillance assets, evidentiary assets, and breach targets simultaneously. The chatbot context is therefore not best analysed pathway-by-pathway in isolation. It is best analysed as a single system in which the same architectural choices feed multiple pathways at once.

The third is the **chilling-effect dimension**. Chapter 1 observed that consumer chatbot use takes place, at scale, in a confided register that closely resembles the speech people address to spouses, close friends, parents, or therapists. The same register makes the speech valuable to users (because conversational AI provides something that other interlocutors cannot or will not) and vulnerable across all four pathways. As the documented cases accumulate, users might progressively come to understand that this speech can be referred to law enforcement, compelled by court order, produced in litigation to which they are not a party, or exposed in a breach, to say nothing of the access by the providers themselves described in detail in Part 1 of this study. At that point the chilling-effect concern moves from the level of individual privacy harm to the level of the technology as a whole: it bears on the willingness of users to address these systems with the same candour, and, through that, on the social value the technology can be expected to deliver.

5.2. Building on existing mechanisms, addressing what is distinctively new

Much of the legal and institutional machinery analysed in this study is not new in its general form. Provider-initiated emergency disclosure under good-faith standards has operated for decades across email, messaging, social-media and cloud-storage providers. Compelled access through subpoenas, court orders, search warrants and preservation requests is the routine output of the criminal-procedure framework of every developed jurisdiction. Civil discovery, including discovery of third-party records, is a well-developed feature of common-law procedural systems. Breach notification, insider-access controls, vendor management, and the broader cybersecurity toolkit have decades of development behind them. The first claim of this study is therefore a conservative one: the chatbot context inherits this existing infrastructure rather than displacing it, and the recommendations that follow build on existing legal and institutional frameworks rather than starting from scratch.

What is new is, in each chapter, a specific combination of features that the established framework was not calibrated for, and that, taken together, produce a setting that adjacent contexts do not. Five features have recurred across the four chapters and are worth restating together.

The first is the *confided register* in which much consumer chatbot use is conducted: users speak to chatbots in ways more reminiscent of conversations with spouses, close friends, parents, or therapists than of public or semi-public posts in social media. **The second** is the *interpretive ambiguity* that surrounds individual utterances in this register, where intent blends with exploration, fiction, role-play, provocation, and venting in ways that resist confident classification, but where the institutional decisions that follow (refusal, suspension, referral, retention, production) are nevertheless made. **The third** is the *scale* at which this register operates: hundreds of millions of users now confide in these systems in conversations that combine intimate disclosure with conversational continuity at automated scale. Social media moderation and large-scale search operate at comparable volumes but principally on material of a different register. **The fourth** is the emergence of *new categories of record* (memory files that build structured longitudinal profiles; agentic action logs that combine natural-language intent with autonomous execution and real-world effect) whose application to the established legal categories of communications and transactional records is not yet stable. **The fifth** is the *absence of the legal-institutional architecture* (privilege, professional duties of care, continuing relational context) that has historically surrounded confided speech in the human settings whose register the chatbot setting borrows.

The novelty is in the conjunction. No individual element of the list above is new on its own. What the chatbot setting brings together, in a single technological category and at unprecedented scale, is a combination of features that does not appear in any of the adjacent settings taken individually. That is what justifies treating the chatbot setting as a distinctive object of analysis, and what shapes the kind of recommendations the study proposes: targeted refinements of an existing framework rather than a wholesale replacement.

5.3. The architectural turn: Meta's 13 May 2026 announcement

A development during the final days of drafting deserves to be registered explicitly in the conclusions. On 13 May 2026, Meta announced *Incognito Chat with Meta AI*, extending its Private Processing architecture from peripheral AI features inside WhatsApp (message summarisation, writing assistance) to the chat conversation itself between a user and Meta AI. Users of WhatsApp can now chat with Meta AI in a mode in which the provider cannot read the conversation, in the closest functional equivalent to the way Meta cannot read WhatsApp messages between two users. The protection is architectural rather than contractual: the chat is processed inside a **Trusted Execution Environment** in which the provider has renounced, by hardware design, the capacity to access content. The announcement is the subject of a dedicated paper with Peter Swire,²¹⁵ and the substantive analysis is not duplicated here.

²¹⁵ Theodore Christakis and Peter Swire, "[The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI](#)", AI-Regulation Papers 2026-05-4, May 2026.

What the announcement contributes to the analytical frame of this study can be stated in three observations.

The first is that the architectural turn from policy-based to architecture-based confidentiality, which Part 1 of this study had proposed under the name of *Sealed Mode*, and which Apple's *Private Cloud Compute* of June 2024 had instantiated in a limited form, has now occurred in a mass-market consumer chatbot context. The intellectual lineage runs from the confidential-computing infrastructure that made the deployment technically possible, through the analytical framework developed in Part 1 that articulated what such a deployment would mean for chatbot confidentiality, to the actual deployment by Meta on 13 May 2026. The first deployment is significant in itself, but the more important question raised by it is whether the other major providers (OpenAI, Anthropic, Google, xAI and others) will follow, at least for some applications (such as ephemeral chats).

The second observation, which the WhatsApp paper develops in detail, is that the Meta architecture resolves much of the privacy exposure documented across the four chapters of this study through a rather direct architectural choice, while opening a new set of tensions that the privacy frame alone cannot answer. On the resolution side: hardware-enforced confidentiality reduces the surface for compelled government access (Chapter 2), narrows the dataset available for discovery (Chapter 3), and shrinks the breach blast radius (Chapter 4). It also closes off the proactive-disclosure pipeline analysed in Chapter 1: the Strasbourg, FSU and Tumbler Ridge episodes all presupposed a provider that could read what users were saying, and that visibility no longer exists in the architecture Meta has now deployed. These are not marginal improvements; the architecture removes much of the retained-data exposure that the four chapters had identified as the dominant privacy risk. On the tension side: the same underlying architectural fact, that the provider has no plaintext access to the conversation, produces a further chain of consequences whose normative direction runs the other way. It removes the "knowledge" element on which the emerging provider-liability claims depend, in the wrongful-death actions arising from Tumbler Ridge and in the Garcia and Raine litigation. The chain is direct: no visibility means no possibility of referral, no possibility of referral means no failure of a duty to refer, and no failure of duty means no liability of the kind those cases have begun to develop. The same fact constrains the provider's capacity for content-based safety monitoring more generally, and opens a new dimension of the long-running Going Dark debate in a setting where the provider is itself one end of the conversation rather than a routing intermediary between human parties. The substantive policy work these new tensions require, namely how to provide adequate safety mitigations in a no-visibility setting, how to allocate responsibility for harms the provider cannot see, and what the Going Dark debate means when the provider is itself a party to the conversation, is the work the architectural turn now puts on the agenda. These are policy questions that follow from architectural confidentiality, not arguments against it.

The third observation is that the architectural turn does not displace the policy layer. By the policy layer this study means everything that is not built into code: not only the contractual, transparency, and accountability commitments that surround chatbot conversations, but the management and operational decisions about retention, internal access, logging, and vendor distribution that determine what data exists in the first place. That layer remains essential and continues to do real work, particularly in the domains the architectural pathway does not cover. What the architectural turn adds, in the specific high-sensitivity domains where it is deployed, is a category of protection that policy alone cannot supply. The two layers operate together; neither makes the other obsolete.

5.4. Open questions

The analytical work on which a stable governance framework can be built is now under way. Several open questions deserve to be flagged for the next phase of research, three of them with particular urgency.

The first concerns whether, on what timeline, and on what conditions other major chatbot providers will follow Meta in deploying TEE-based architectures for confidential chatbot pathways. The technical building blocks are now in two production deployments (Apple's *Private Cloud Compute* and Meta's *Private Processing*); the engineering tractability at consumer scale is no longer in question. What

remains uncertain is whether the standalone providers (OpenAI, Anthropic, Google, xAI, Mistral...) will choose to do so, for which features, on what timeline, and under what user-facing framing. A related question is whether confidential-memory architectures can in time be developed that preserve the persistence and personalisation features users currently value, without recreating the centralised surveillance surface that statelessness avoids. Beyond the six providers most directly addressed in this study, dedicated analysis will also be needed both for emerging European entrants such as Mistral, whose deployment choices and surrounding legal environment differ from those of the US-based providers covered here, and for the position under Chinese national-security and intelligence-cooperation law that frames DeepSeek's consumer chatbot.

The second question concerns whether Meta's architectural turn will face the same pressures from law-enforcement agencies, online-safety regulators, and child-safety organisations that the user-to-user E2EE debate has faced for two decades. The arguments deployed against the *Incognito Chat with Meta AI* architecture will be familiar from prior rounds of the encryption debate: the importance of provider-side visibility for the investigation of grave offences, the responsibility of providers for harms produced by their own systems, and the asymmetry between law-abiding users and motivated bad actors who retain access to alternative channels. A closely related question is whether the safety mitigations that operate inside the architecture (model-level refusal, in-enclave classifier moderation, crisis-resource redirection for self-harm queries, and abuse-pattern throttling for repeated jailbreak attempts) will be judged adequate by regulators, civil-society organisations, and the public when individual harmful outputs inevitably surface in reporting or in litigation. The architectural turn will be tested not only on its abstract merits but on how it performs in cases the public can see.

The third concerns how the legal frameworks examined in the four chapters will respond to the architectural turn. The empirical case for a limited AI privilege gains analytical support from architectures of the kind Meta has now deployed (Chapter 3). The case for distinguishing memory and agentic action logs from ordinary chat transcripts gains it as well (Chapter 2). The constitutional debate over reverse-prompt warrants, particularly in their potential bulk form, takes a different shape when the corpus to be searched is architecturally inaccessible to the provider (Chapter 2). And the proactive-disclosure framework analysed in Chapter 1 faces the opposite pressure: the referral and safety-monitoring expectations now forming around it all presuppose a provider that can read what users write, which is precisely the assumption the Encrypted Chats in WhatsApp deployment removes. Stable doctrinal answers to these questions are unlikely to emerge before several further years of litigation and regulatory practice.

This study should therefore be read as it was always intended: a first map of a rapidly moving field, offered in the spirit of opening a debate rather than closing it, and as a contribution to the work that practitioners, scholars and institutions on both sides of the Atlantic are already starting to undertake.

5.5. Consolidated Tentative Recommendations

The recommendations that follow consolidate, in a single condensed form, the tentative recommendations set out at the end of each of the four chapters. They are addressed in the first instance to providers, and within a provider to product, engineering, security, and senior management as much as to legal and policy teams, and secondarily to regulators and oversight bodies on both sides of the Atlantic, as well as to the academic and practitioner communities whose engagement will be needed for the substantive design work. The numbering follows the four-chapter sequence: 1–5 on proactive corporate disclosure, 6–8 on compelled government access, 9–14 on discovery and evidence, and 15–19 on cyberthreats, insider risks and data breaches. Each chapter's rows are shaded in a different colour so the four groups can be visually distinguished within the consolidated table.

#	Recommendation
Proactive Corporate Disclosure (Chapter 1)	
1	Aggregate transparency reporting on provider-initiated emergency referrals. Major chatbot providers should publish aggregate data on their provider-initiated emergency referrals to law enforcement, including volume, the categories of concern that triggered them, and the jurisdictional distribution of receiving authorities. The reporting model is the proactive CSAM reporting these same providers already operate at scale.
2	Audit mechanisms for the referral pipeline. Providers and regulators should consider periodic audits of the criteria that determine what enters the referral pipeline, building on existing accountability structures (§ 2702 good-faith requirements, equivalent EU provisions, civil-rights litigation, press scrutiny). Candidate models include audits of the criteria themselves, reviewable internal records of how they are applied, and, where legally permissible, post-hoc user notification.
3	Retention: distinguish preservation, user-elected storage, and provider-side retention. These three concepts are sometimes conflated and warrant separate treatment. Preservation orders operate within an established legal framework; user-elected storage should be governed by user control; provider-side retention beyond what the user has elected creates a record that serves both legitimate post-incident cooperation and purposes the user did not elect, and that tension warrants further analytical work.
4	Cryptographic hardening calibrated to the deployment environment. The level of cryptographic protection appropriate for a chatbot depends on the environment in which it operates (standalone consumer chatbot, chatbot embedded in an end-to-end encrypted messaging service, Sealed Mode), because that environment carries prior commitments to users that the chatbot inherits. Any policy discussion of cryptographic protection should begin by identifying which environment is in play, rather than treating all chatbot conversations as a single category.
5	User-facing transparency about the nature of the chatbot relationship. Providers and regulators should develop chatbot-specific transparency disclosures that meaningfully communicate the operating conditions of the system at the point of use, since chatbot speech is conducted in the register of confidant relationships but receives almost none of the legal-institutional protections that surround such speech in adjacent human contexts.
Compelled Government Access (Chapter 2)	
6	Improve transparency reporting on compelled access to chatbot data. Providers should make such access as legible as the law allows and, where permitted, report separately the two categories

#	Recommendation
	peculiar to this setting: reverse-prompt demands and demands targeting memory or personalisation features.
7	Confirm publicly that the commitment to resist overbroad demands reaches bulk reverse searches. Providers' existing commitments to resist overbroad demands already cover most of the new categories; the one worth stating specifically is the most novel and most dangerous, a bulk or content-pattern search across the prompt database with no nexus to an identified user. Saying plainly that such demands will be resisted, as an illustration rather than a narrowing of the general commitment, gives courts and users a concrete expectation where the posture is least settled.
8	Treat memory and agentic action logs as distinct legal-process categories. Providers already separate content from transactional data as a matter of routine; the narrower step recommended here is that memory profiles and agentic action logs be named as distinct record types in legal-process review, transparency reporting, retention design and user-facing explanations. Memory files are structured longitudinal profiles, and agentic logs bind natural-language intent, autonomous execution and real-world effect in a single record, so treating them explicitly makes visible a distinction the established framework otherwise handles only implicitly.
Discovery and Evidence (Chapter 3)	
9	Design products with discovery exposure as an explicit consideration. The NYT discovery precedent shows that stored chatbot conversations are vulnerable to large-scale production orders. Providers should treat architectural choices (retention limits, memory and conversation-history design, Sealed Mode pathways for the most sensitive categories) as decisions with discovery implications, not only privacy ones. Reducing retention where the user so desires, and architecting verifiable Sealed Mode for the most sensitive categories, narrows the surface available for litigation capture.
10	Pursue a limited AI privilege for the most sensitive chatbot conversations, supported by architecture rather than by terms of service alone. No general privilege protects chatbot conversations from discovery under current US law, and recognising one would require legislative, rule-making, or judicial action. Each route requires objective indicators of confidentiality at the time of the exchange, which ordinary terms of service cannot supply but verified Sealed Mode conditions could. Providers should build the architectural foundations on which a privilege claim can rest.
11	Resolve the work-product question raised by <i>Heppner</i> and <i>Warner</i>. <i>Heppner</i> treated a litigant's AI-assisted preparation as disclosed to a third party and so unprotected; <i>Warner</i> treated the chatbot as a tool the litigant used and left work-product protection intact. Courts, and in time the rule-makers, should clarify that using a general-purpose AI tool to assist one's own litigation preparation does not by itself waive work-product protection, while leaving the ordinary limits of the doctrine, including the crime-fraud exception, intact.
12	Develop chatbot-specific proportionality guidance for civil litigation. Federal Rule of Civil Procedure 26 requires proportionality, but no stable guidance yet exists on how it should be assessed where the discovery target is the bulk conversational record of millions of non-party users. The substantial US institutional work already under way on generative AI and eDiscovery, notably at The Sedona Conference, would be a natural setting for the situation foregrounded by the New York Times litigation to be addressed.
13	Strengthen de-identification and govern derivative materials. De-identification of chatbot conversations is harder than for conventional records because the substance of a conversation can itself be identifying, as European data protection law has long recognised in distinguishing pseudonymisation from anonymisation. Providers should develop chatbot-specific protocols addressing rare fact combinations and references to unique personal circumstances. Protective

#	Recommendation
	orders for large-scale production should also address derivative materials (excerpts, datasets, summaries, embeddings), limiting their use to the specific proceeding.
14	Warn users, especially in sensitive contexts, that their conversations may be discoverable. Sam Altman's July 2025 acknowledgment that ChatGPT conversations lack legal confidentiality was important but insufficient. Providers should inform users clearly, and in plainly visible form, that ordinary consumer chatbot conversations may be subject to lawful production in litigation, particularly in health, mental health, legal, financial, and other high-sensitivity contexts. Such warnings should be paired with a genuine protected alternative, such as Sealed Mode, where available.
Cyberthreats, Insider Risks and Data Breaches (Chapter 4)	
15	Treat chatbot logs as high-sensitivity data by default. Providers should classify consumer chatbot logs, memory files, uploaded documents, and agentic action logs as high-sensitivity datasets for security purposes, even when no special-category data has been identified. The service is designed to receive intimate, legally sensitive, medical, financial, professional, and emotional disclosures at scale; security controls should be calibrated to the data the system predictably receives, not to the lowest-sensitivity conversation it might contain.
16	Reduce the breach blast radius through minimisation, segmentation, and Sealed Mode as a security control. The safest chatbot log is the one that does not exist or does not persist. Providers should minimise default retention, separate chat history from memory profiles, isolate uploaded files, and avoid unnecessary replication into analytics, support, or model-evaluation systems, including flows of conversation metadata to third parties. The Sealed Mode proposed in Part 1 should be designed as a security architecture as well as a privacy feature, with strict retention, limited human access, and stronger cryptographic protections.
17	Harden access controls across insiders, vendors, and AI agents. The same least-privilege logic applies across all three categories. For insiders and vendors: just-in-time approval, segregation of duties, logging of access to conversation datasets, and contractual controls for third parties. For AI agents acting on the user's behalf: short-lived tokens, user confirmation for high-risk actions, and strong controls against indirect prompt injection. Agents that can read email, files, health records, or enterprise repositories should be treated as security-sensitive applications.
18	Design sharing interfaces and breach response with chatbot-data sensitivity in mind. The ChatGPT indexing episode and the Meta AI Discover controversy showed that sharing interfaces designed for ordinary content can produce serious exposure when applied to confidential conversations. Sharing features should default to private, require clear friction for public sharing, and avoid indexing. Where chatbot data is exposed, breach assessment should look at substance and context, not only identifiers; a technically narrow breach may still be personally devastating.
19	Publish AI-specific security and breach transparency. Transparency reports should not be limited to government requests and CSAM reports. Providers should publish aggregate information about material security incidents affecting chatbot data, to the extent compatible with security and legal obligations. Users and regulators need to understand not only how often the state asks for chatbot logs, but also how often the systems holding those logs fail.