AI-Regulation Papers 2026-03-2

You Trust Your Chatbot With Everything. Should You?

Part 1: How The Controller Uses Your Chat Data

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.

Acknowledgments and AI use:
The author would like to thank all colleagues who agreed to peer review this article on short notice. The author's warmest thanks go to Peter Swire and Joe Jones for their invaluable suggestions on an earlier draft. Many thanks also to Claude Castelluccia,Eleni Kosta, Ken Propp, Mikołaj Barczentewicz, Eleni Polymenopoulou, Thodoros Karathanasis, Shadée Pinto, and Pankaj Raj. Any remaining errors are the author's own.
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.



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 1: How The Controller Uses Your Chat Data, AI Regulation Papers, 26-03-2, AI-Regulation.com, March 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).

You Trust Your Chatbot With Everything. Should You?

Part I: How the Controller Uses Your Chat Data

EXECUTIVE SUMMARY

Chatbots have become the world's most trusted strangers. Every day, hundreds of millions of people confide health symptoms, legal strategies, financial anxieties, relationship crises, and moments of acute emotional distress to systems that feel private but are not governed by anything resembling professional secrecy. The interface invites intimacy; the fine print reserves broad rights most users will never read. **This study is the first comprehensive academic attempt to map the internal privacy boundary of consumer chatbot conversations**: how providers handle the data users entrust to them, where the protections fall short of what the interface invites users to expect, and what constraint-based alternatives could look like.

Through a comparative policy-and-interface analysis of five major services (ChatGPT, Gemini, Claude, Grok, DeepSeek), including four detailed comparative tables, Part I examines the *internal* boundary: how the provider may reuse conversations for training, review them through human annotators, monetise them through advertising, and circulate them across vendors and connected services. Part II, forthcoming, examines the *external* boundary: civil discovery, government-compelled access, and cybersecurity breaches.

These recommendations are not a European compliance checklist. They are conditions for sustainable trust in a market where trust is becoming a differentiator and, increasingly, a competitive moat. Uncertainty about confidentiality produces concrete harm: users self-censor, avoid the highest-value use cases, or over-disclose under false assumptions. Even users who profess indifference to privacy care about reputational harm, professional consequences, or the fear that a personal vulnerability might resurface in an unexpected context. The recommendations of the present paper need to be tested, but they intend to preserve the utility, safety monitoring, and commercial viability of chatbots.

What the study found. The findings do not reveal a landscape of abuse, but they do reveal a landscape of structural opacity:

Training is the norm. Every major provider now trains on consumer chats by default. One forces users into a trade-off between chat continuity and broader data reuse. At least two others allow a single thumbs-up/down click to silently override your opt-out for an entire conversation. Another one ties the training choice directly to retention: opting in extends storage from 30 days to five years, so that a user who agrees to "help improve" the model is also accepting a significantly longer backend retention period, although user-initiated deletion overrides it.

Human review is structural. Every provider reserves the right for humans to read conversations. Only one (Gemini) warns users in the interface: "*don't enter confidential information or any data you wouldn't want a reviewer to see*". Reviewed chats can be retained for up to three years after the user deletes them; safety scores may persist for seven years.

Advertising has entered the chat. In February 2026, OpenAI began testing ads in the United States in ChatGPT for logged-in adult users on the Free and Go subscription tiers. Ad personalisation is enabled by default and, where memory is also on, draws on past chats and stored memories to select ads. Other providers position their consumer chat experiences as ad-free; the question is whether this posture will remain durable as business models evolve and monetisation incentives grow.

"No sale" does not resolve the full transparency question. All five providers make some version of a "no sale" commitment, and these commitments are genuine: user data is not transferred for third parties' independent commercial use. But operational sharing with processors and vendors, though legally constrained under the GDPR's controller–processor framework and often undertaken for privacy-protective purposes such as safety auditing, remains under-disclosed. Users typically cannot assess who may access their conversations or

under what safeguards. For two providers, the chatbot is part of a broader ecosystem where cross-service data flows are a feature, not an exception.

A centrepiece proposal: Sealed Mode. The study's most original recommendation addresses a gap that no existing consumer feature fills. Millions of people already use chatbots as substitutes for protected relationships, discussing symptoms, medication, and grief in an environment that feels confidential but carries no meaningful insider-access constraints. Warning labels alone cannot fix this: telling users not to share sensitive information while the product design keeps inviting disclosure is, as the EDPB has noted, an incomplete response that risks shifting responsibility onto users.

The study proposes **Sealed Mode**: a clearly labelled consumer pathway (e.g. "Health & Wellbeing – Sealed Mode") where the default architecture materially constrains reuse and insider access, combining no training, no advertising, siloed personalisation, strict retention, minimised human review, and cryptographic hardening that moves the privacy boundary from *promise-based* to *constraint-based.* This could be feasible at scale. The paper does not prescribe a single architecture. It establishes that the most sensitive conversations deserve protections commensurate with the trust users place in them, and calls on stakeholders to treat this as a priority design problem.

The table below consolidates all ten recommendations. Three imperatives run through them: **genuine transparency** at the point of disclosure; **comprehensible, actionable controls** for ordinary users; and **privacy by design,** because policy promises not embedded in product architecture will not sustain the trust that billions of users place in these systems every day.

Consolidated Recommendations from Part I

#	Recommendation
1	**Decouple history, retention, and training.** Allow full conversation history while keeping training disabled. Retention should not vary with the training choice.
2	**Make opt-out feedback-proof.** Separate training, analytics, and safety in controls. A feedback click must not silently override an account-level opt-out.
3	**Regurgitation-aware safeguards.** Publish mitigation measures and residual risk. Offer, to the extent possible, high-stakes users a no-training mode with strict, verifiable retention.
4	**Standardised interface disclosure of human review.** Clear notice in the interface that chats may be reviewed by humans, with retention posture for reviewed chats.
5	**Event-based transparency.** Where feasible, notify users when a conversation is escalated for review, or provide an audit log of access categories.
6	**Sealed Mode for high-stakes topics.** A clearly labelled sealed lane (e.g. "Health & Wellbeing – Sealed Mode") with no training, no ads, siloed personalisation, strict retention, minimised human access, and cryptographic hardening.
7	**Regulatory-first approach to conversational advertising.** Map all applicable rules before deployment. If a solution cannot safeguard both privacy and user trust, do not deploy ads at all.
8	**Recipient transparency.** Describe which vendors and teams access plaintext chat content and which receive only derived or redacted data.
9	**Ecosystem boundary notices.** When a chatbot shares data with another service, disclose it at the moment it occurs and provide a control to disable it.
10	**Separate "operations" from "improvement".** Separate analytics, safety, and training in user controls and disclosures.

Part II will examine the boundary beyond the provider: civil discovery and litigation holds, government-compelled access, and cybersecurity breaches amplified by the retention choices documented here.

You Trust Your Chatbot With Everything. Should You?

Part I: How the Controller Uses Your Chat Data

By Theodore Christakis

INTRODUCTION

The Problem

A simple moment now recurs hundreds of millions of times a day. A user opens ChatGPT, Gemini, Claude, or another consumer chatbot, pauses, and types something they would hesitate to say aloud to a colleague, a doctor's receptionist, or sometimes even a friend. The hesitation is not about whether the model can provide a useful answer. It is about whether the conversation is truly safe to have.

The questions are immediate and intensely practical. Can someone at the company read this exchange? Is my prompt treated as confidential, or as raw material for model improvement? If I disclose a medical concern, a family dispute, or a draft legal strategy, could it be reviewed by human annotators, retained for extended periods, or fed into training pipelines? If I am a researcher testing an unpublished argument, could fragments resurface in responses served to others? If I seek legal or financial guidance, could this conversation become a record that later works against me? And as consumer AI matures into a business, another worry follows: am I being profiled, categorized, or monetized through what I confide?

Far from being hypothetical, these questions arise at scale. A growing body of empirical research confirms that users routinely share highly sensitive personal information with chatbots, including health data, relationship details, and professional secrets, often driven by the perceived anonymity and non-judgmental quality of the interaction.[1] Yet studies also reveal a striking disconnect: many users fundamentally misunderstand how their data is handled, and even those who express privacy concerns do not consistently protect themselves when chatting.[2] Meanwhile, a recent Stanford HAI study examining the privacy policies of six major U.S. chatbot developers found that every one of them trains

[1] Research on self-disclosure to conversational AI consistently finds that users share intimate information with chatbots as readily as (or more readily than) with human interlocutors, driven by perceived anonymity and reduced fear of judgment. Key references include: Emmelyn Croes et al., "Digital Confessions: The Willingness to Disclose Intimate Information to a Chatbot and its Impact on Emotional Well-Being" Interacting with Computers 36, no. 5 (2024): 279–292; Niloofar Mireshghallah et al., "Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild" COLM 2024; Hao-Ping Lee et al., "Understanding Users' Privacy Reasoning and Behaviors During Chatbot Use to Support Meaningful Agency in Privacy" arXiv:2601.18125 (2025).

[2] On the "privacy paradox" in the chatbot context — the gap between users' stated privacy concerns and their actual disclosure behavior — see Julia Ive et al., "Privacy-Preserving Behaviour of Chatbot Users: Steering Through Trust Dynamics" arXiv:2411.17589 (2024, revised 2025), finding that 76% of participants lacked a basic understanding of privacy risks in chatbot interactions. See also Daniel Solove, "The Myth of the Privacy Paradox" George Washington Law Review 89 (2021): 1–51.

on consumer chat data by default, that some retain this data indefinitely, and that opt-out mechanisms vary widely in visibility and effectiveness.[3]

The central problem is structural. Chatbot interfaces are designed to feel personal and low-friction; they invite intimacy. But the legal terms and product settings that govern what happens to the resulting conversations are dense, scattered across multiple documents, and written in language that most users will never parse at the moment of disclosure.[4] The gap between the intimacy that the interface encourages and the uses that providers reserve for themselves is what makes the privacy question in consumer AI distinct from older debates about data collection by search engines or social-media platforms.[5]

State of the Art

Over the past two years, both academic scholarship and regulators have understandably focused on the upstream question: whether and on what legal basis general-purpose AI models are trained with personal data harvested on the internet, and what governance constraints should apply to that training. That focus has generated important work on lawful bases, transparency, purpose limitation, and the special risks created by web-scale data collection.[6] Yet it has also left largely under-examined what may be the elephant in the room: the downstream privacy of everyday, often highly intimate conversations between consumers and chatbots. Hundreds of millions of users in Europe and worldwide now use these systems to discuss health concerns, family conflicts, professional secrets, and legal strategies, but there is still very little detailed academic analysis of how such interactions are actually protected in practice, and almost no work that examines, in a single holistic framework, the combined effect of policies, default settings, interface warnings, opt-outs, retention rules, human review pathways, and emerging monetization incentives, as well as the external exposure vectors that can defeat confidentiality even when providers intend to preserve it, including civil discovery and litigation holds, legally compelled government access, and breach-driven disclosure risks that are amplified by retention and logging choices.

This series is designed to fill that gap by producing a practical map of the confidentiality boundary as it exists for ordinary users, and by translating that map into concrete privacy-by-design recommendations that can strengthen trust without undermining safety.

---

[3] Jennifer King et al., "User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies" Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8, no. 2 (2025): 1465–1477. See also the Stanford HAI overview: Be Careful What You Tell Your AI Chatbot. The study examines privacy policies of Amazon (Nova), Anthropic (Claude), Google (Gemini), Meta (Meta AI), Microsoft (Copilot), and OpenAI (ChatGPT), finding that all six train on user chat data by default, some retain data indefinitely, and opt-out mechanisms vary in clarity and effectiveness.

[4] OpenAI, "How Your Data Is Used to Improve Model Performance" (policy explainer). This page details how consumer chat inputs may be used for training and illustrates the type of disclosure that users would need to locate and parse to understand the controller's internal practices.

[5] Jennifer King, "Safeguarding Data Privacy and Well-Being for AI Chatbot Users" Testimony before the U.S. House Committee on Energy and Commerce, Subcommittee on Oversight and Investigations, 18 November 2025. King identifies a "core misalignment between how chatbots are designed and how the public uses them," noting that "general-purpose chatbots are designed to maximize consumer engagement and have no fiduciary or professional responsibility to put the well-being of their users above their business model."

[6] Among academic works, for instance: Araz Taeihagh, "Governance of Generative AI", Policy and Society 44 (2025): 1–22; Hannah Ruschemeier, "Generative AI and data protection", Cambridge Forum on AI: Law and Governance. 1 (2025); Wenlong Li, Yueming Zhang, Qingqing Zheng, Aolan Li, "How the Legal Basis for AI Training is Framed in Data Protection Guidelines and Interventions: Comparative Perspectives and the Prospect of Global Convergence", *International Data Privacy Law*, (2026).

The two parts of the series correspond to two privacy boundaries.

The present Part I examines what happens *inside the controller*: how the provider itself may reuse chat interactions[7] for future training, improvement, safety review, commercialization and other purposes.

Part II, to be published shortly, examines what happens *beyond the controller*: when and how chat data may be disclosed or accessed through civil discovery, government demands, and broader exposure vectors, including retention-driven vulnerability and breach risk.

Together, the two parts aim to provide a comprehensive legal and policy map of the lifecycle of a consumer chatbot interaction, from the moment of input to the moment of potential external exposure.

A Comparative Policy-and-Interface Analysis

The objective of the series is to map the true boundaries of consumer chatbot privacy through a comparative policy-and-interface analysis.

The series does not speculate about undisclosed internal practices, nor does it claim to audit how providers actually implement data handling inside their systems. It focuses instead on what can be responsibly established from the public record, by triangulating the main sources through which providers communicate their privacy posture to ordinary users: the privacy policy and consumer terms of use; official help-center and privacy-hub documentation that explains controls, modes, and retention; the disclosures embedded directly in the user interface (warnings, toggles, and settings screens); and other public communications that materially shape privacy expectations, such as product blog posts and policy explainers announcing new features (for example, temporary chat modes or advertising-related controls). This broader documentation set is essential because the most operationally significant details are often not in the privacy policy itself, but in help-center pages and in-product prompts encountered only at the moment users decide what to disclose.

A particular focus of this series is the recent spread of **"temporary" or "ephemeral" chat modes** (also described as "Incognito" or "Private" chats). These modes are increasingly presented as a practical way for users to reduce persistence and downstream reuse. They generally aim to keep a conversation out of ordinary chat history, limit or disable personalization and "memory," and exclude the session from model training, while still allowing some short-term retention and access for service provision, security, and abuse prevention. Because these modes can materially reshape the confidentiality boundary that users perceive at the moment of disclosure, this study treats them as a core part of

[7] In this series, "interaction data" is used in a broad, practical sense to capture the full set of information generated through consumer chatbot use. It includes (i) user prompts and the model's outputs, including follow-up turns within a thread; (ii) any files, images, or other materials uploaded or pasted into the interface; (iii) voice inputs and audio outputs where voice modes are used; (iv) technical and account metadata associated with the interaction, such as device and browser/app identifiers, IP address, coarse location signals, timestamps, account identifiers, and usage logs; (v) "memory" and personalization features that store user-provided facts or inferred preferences to shape future responses; and (vi) safety and integrity signals associated with moderation or abuse detection (for example, flags, classifier outputs, or related risk indicators). Unless otherwise specified, the comparative analysis focuses primarily on interaction content (prompts, outputs, and user-provided attachments) and the closely related metadata that providers explicitly tie to retention, review, personalization, training, monetization, or disclosure pathways.

the consumer privacy architecture and analyzes them as disclosed in both policies and interface-level documentation.[8]

Where these sources remain incomplete, inconsistent, or ambiguous, the analysis does not fill gaps by assumption. It flags the ambiguity, and treats that opacity as a finding in its own right, because uncertainty about the confidentiality boundary predictably encourages the very over-disclosure that the research literature documents.[9] All documentation and UI captures were collected between 1 and 23 February 2026.

Scope and Methodology

The scope is intentionally limited to five widely used and "consumer, general purpose, conversation-first" chatbot services: ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), Grok (xAI/X), and DeepSeek Chat (DeepSeek). This choice is both methodological and practical.

Methodologically, these services sit at the center of everyday, long-form chatbot use worldwide, and (according to the latest available global web traffic benchmarks[10]) they are among the most visited consumer chatbot destinations. I opted for four major US-based providers and one China-based provider to enable a meaningful cross-jurisdictional comparison.

Practically, a truly comprehensive study covering every consumer chatbot would be unmanageable: privacy policies, settings, and interface disclosures change frequently; important controls are sometimes scattered across multiple documents; and accurate comparison requires close, repeated verification of the most current sources.

I therefore chose depth and precision over breadth. I also excluded several other widely discussed systems because they do not fit the "standalone consumer chatbot" model that this series targets. Some are primarily embedded into broader software or social ecosystems, where the relevant privacy boundary is not only the chatbot's policy but also the surrounding platform's data flows and defaults, for example Microsoft Copilot (deeply integrated into Windows and Microsoft 365) or Meta AI (embedded across WhatsApp, Instagram, Messenger, and Facebook). Others are retrieval-first "answer engines" whose primary consumer use case and disclosure logic differs materially from intimate, conversation-style assistance (for example, Perplexity).

Enterprise offerings, negotiated business contracts, and API access are excluded from the scope of this series because they typically involve materially different commitments, technical configurations, and governance assumptions than consumer chatbots. The analysis therefore does not cover enterprise or workplace plans; developer/API usage; "bring your own key" or custom-hosted configurations; or sector-specific deployments (for example, in healthcare, legal services, or other regulated professional environments) where confidentiality may be shaped by contract, professional secrecy, or domain-specific compliance programs. The focus is intentionally

[8] For example, ChatGPT offers a "Temporary Chat" mode that many ordinary users overlook. It can be enabled directly from the chat interface (via the Temporary Chat option at the top of a new conversation), and is designed so that the chat does not appear in history and does not create or update "memory." See help.openai.com.

[9] Emel Gumusel, Kevin Z. Zhou & Madelyn R. Sanfilippo, "A Literature Review of User Privacy Concerns in Conversational Chatbots: A Social Informatics Approach" Journal of the Association for Information Science and Technology 75 (2024): 1377–1393.

[10] The latest publicly reported Similarweb "Global AI Tracker" snapshot of worldwide GenAI chatbot website traffic share (January 2, 2026), places ChatGPT (64.5%) and Gemini (21.5%) far ahead, followed by DeepSeek (3.7%), Grok (3.4%), Perplexity (2.0%), Claude (2.0%), and Copilot (1.1%).

confined to the standard consumer experience: ordinary adult users interacting through free access or consumer subscriptions in the public, first-party interfaces.

This study focuses on adult interactions with consumer chatbots. It **does not examine minors' use (under 18)**, which raises distinct and often heightened legal and policy issues, including under EU law. Providers have increasingly introduced age-gating mechanisms and child or teen specific safeguards, and in some cases separate supervised-account regimes, parental controls and dedicated information pages. Those design choices, and their effectiveness, merit a dedicated analysis of their own. Accordingly, the present series brackets children's privacy and safety questions and concentrates on the privacy boundary faced by adults using the standard consumer interfaces.

The precise scope of "consumer access"[11] for each provider is set out in a detailed footnote.[12] In each case, the decisive criterion is whether the plan or tier is marketed directly to individual consumers and governed by the provider's standard consumer-facing policies, rather than by an enterprise or negotiated agreement. This distinction matters practically because, as King et al. have documented, the incorporation of user data into model training by default is a practice that is largely unique to consumer products; enterprise users' data is typically excluded from training unless they explicitly opt in.[13]

The Critical Issue of "Voice Modes"

This study focuses on text-based interactions. It **does not** systematically examine **voice modes**, which are now available across several of the five services studied (ChatGPT Advanced Voice Mode, Gemini Live, Grok voice). Voice interaction raises additional and, in some respects, distinct privacy concerns that deserve dedicated analysis. Three deserve brief mention here.

First, voice lowers the threshold of disclosure. There is little empirical research directly comparing self-disclosure in voice and text chatbot interactions, but it is reasonable to expect that voice lowers the threshold in at least some respects: spoken conversation is faster, more spontaneous, and harder to self-edit than typing, and voice modes are specifically designed to respond with emotional nuance and conversational warmth that may encourage users to confide more freely than they would through a keyboard. Voice modes are designed to maintain conversational flow, respond with emotional nuance, and make the interaction feel closer to speaking with a trusted person than to using a digital

[11] For purposes of scope and comparability, "consumer use" in this series refers to interactions conducted through the providers' first-party consumer interfaces (web applications and official mobile apps on iOS/Android). It excludes third-party clients, unofficial wrappers, and browser extensions, and it also excludes integrations embedded into operating systems, productivity suites, or other applications (for example, OS-level assistants or in-app side panels) unless the chatbot is presented as the primary, standalone interface. This delimitation is important because the applicable disclosures and controls, as well as retention behavior and warning prompts, may vary across surfaces, and a policy-and-interface analysis must compare like with like.

[12] For ChatGPT, the comparison covers the Free tier and consumer subscription tiers described in OpenAI's pricing page (including, at the time of writing, Go, Plus, and Pro), while setting aside Business and Enterprise. See also OpenAI, "Introducing ChatGPT Go". For Gemini, the analysis covers Gemini Apps and the consumer subscription bundles extending access through Google One and related consumer plans, as described in the Gemini Apps Privacy Hub. For Claude, the analysis covers Anthropic's consumer-facing plans (Free and paid tiers, per Claude pricing and Anthropic's Privacy Policy) and excludes Team and Enterprise deployments. For Grok, the analysis covers the consumer experience delivered through X and xAI's consumer-facing documentation (see X Help Center: About Grok). For DeepSeek Chat, the analysis relies on primary disclosures on the DeepSeek site; where subscription structures and controls are not clearly documented, they are not inferred.

[13] Op.cit.

tool. If text-based chatbots already encourage over-disclosure, voice modes are likely to intensify it.

Second, voice carries data that text does not. An audio recording or a natively multimodal voice session encodes not only the semantic content of what the user says, but also paralinguistic signals: stress, hesitation, emotional state, speech patterns, and potentially identifiable vocal characteristics. Under EU data protection law, voice data can constitute biometric data when processed for the purpose of uniquely identifying a natural person (Article 9 GDPR), and even when it does not reach that threshold, it remains personal data whose processing requires a lawful basis and appropriate safeguards. The retention and human review questions examined in this study therefore take on an additional dimension for voice: are audio recordings retained, or only transcripts?[14] Can human reviewers listen to the original audio?

Third, voice blurs the boundary between foreground and background disclosure. A text prompt contains only what the user deliberately types. A voice session can capture ambient sound, background conversations, and statements by third parties who have not consented to any processing. Reports have documented instances where ChatGPT's voice mode continues processing audio after the user stops speaking, capturing exchanges that were never intended as inputs. This raises not only data protection concerns for the user, but third-party privacy issues that have no equivalent in the text-based context.[15] These three dimensions, lowered inhibition, richer personal data, and ambient capture, suggest that voice modes amplify every privacy risk documented in the present study. A comprehensive analysis of voice-specific privacy boundaries, including retention, human review of audio, biometric data processing, and third-party exposure, is a priority for future work.

Framing, Trust, and Why Privacy-by-Design Matters Beyond Europe

This article is written from a particular vantage point: that of a European lawyer working in an environment shaped by heightened data protection norms and an unusually strong cultural and regulatory attention to privacy. It would be artificial, and arguably misleading, to pretend otherwise. Yet the purpose of adopting this prism is not to export the GDPR as a universal template, nor to argue that the world should converge on European legal categories of individual control. The central object of this paper is more elementary and, in that sense, more widely shared: it is the problem of user trust in a technology that increasingly invites disclosure at an intimate scale. Whether a user lives in Paris, New Delhi, São Paulo, or New York, the same practical questions arise when they confide to a chatbot what they would not readily share elsewhere: who might read this, what will happen to it tomorrow, and can it be used against me. The paper therefore treats

[14] Google's Gemini Apps Privacy Hub describes a separate control for whether the audio/video/screens shared through Gemini Live or the mic may be used to improve Google services, and states that Gemini Live recordings are not used for improvement by default. At the same time, it explains that Live chat transcripts and other items shared with Live remain covered by the broader "Keep Activity" setting, a relationship that may not be immediately intuitive to users. OpenAI's Voice Mode documentation states that audio and video clips are retained as long as the chat remains in the user's history; once a chat is deleted, the associated clips are scheduled for deletion within 30 days, subject to exceptions for security or legal retention

[15] See Basil AI, "ChatGPT Voice Mode Caught Recording Conversations Without User Consent", 4 December 2025. On the third-party dimension, Google's Gemini Privacy Hub now explicitly advises users to "ask permission before recording" others in a Live chat: Gemini Apps Privacy Hub. .

transparency and privacy-by-design not as compliance slogans, but as conditions of predictable and sustainable trust in consumer AI systems.

This trust-based framing also addresses a natural skepticism: why should readers accept the paper's normative principles, and why should those who are not especially privacy-conscious care. The first answer is that trust is not a "privacy enthusiast" preference. It is a precondition for adoption in precisely the domains where chatbots are most valuable and most sensitive: health, mental well-being, education, employment, family conflicts, creative work, and legal and financial planning. In these contexts, uncertainty about confidentiality does not merely create abstract privacy risk. It produces concrete behavioral distortions: users self-censor, avoid high-value use cases, withhold key details, or, conversely, over-disclose under false assumptions.[16] Either outcome degrades utility, harms decision-making, and increases the likelihood of downstream disputes, complaints, and regulatory backlash. Even users who profess indifference to privacy often care deeply about a narrower set of concerns that are functionally privacy problems: reputational harm, professional consequences, identity theft, extortion, doxxing, or the fear that a personal vulnerability might be resurfaced in an unexpected context. The boundary question is therefore not "how much privacy should the law require," but "how predictable and defensible is the confidentiality bargain that the product implicitly invites."

For providers, the strategic implication is straightforward. In a competitive market where model capabilities are converging, trust is not a soft value. It is a differentiator and, increasingly, a moat. A chatbot that users perceive as a safe place to think, draft, explore, and confess will capture higher-intensity use, stronger retention, and broader institutional acceptance.[17] Conversely, even a perception of creeping monetization, opaque reuse, or uncontrolled human access can trigger rapid loss of confidence, litigation exposure, and regulatory scrutiny. This paper's recommendations are therefore not offered as a checklist for legal compliance in any particular jurisdiction. They are offered as practical design and governance moves aimed at protecting the trust relationship itself, in a way that should be intelligible and valuable across jurisdictions.

[16] These two responses, over-disclosure and self-censorship, may appear contradictory, and a reader familiar with the privacy paradox literature will recognise the tension. But the paradox dissolves once both are understood as symptoms of the same underlying condition: an opaque confidentiality boundary. Users who cannot reliably assess how their conversations will be handled will sometimes disclose too much (when the interface feels safe) and sometimes withhold too much (when doubt surfaces). The problem is not that users are irrational; it is that the product gives them insufficient basis for rational calibration. Both pathologies are therefore evidence for, not against, the need for the transparency and design reforms this study proposes.

[17] A growing technical constituency is choosing to run open-weight models (such as Meta's Llama, Mistral, or DeepSeek's open-weight releases) locally, on their own hardware. Local deployment eliminates the entire internal boundary that this study maps: there is no provider-side training on conversations, no human review, no vendor sharing, no advertising, and no retention beyond what the user controls on their own machine. That said, local deployment is not without trade-offs. Locally hosted models are typically smaller and less capable than their cloud-served counterparts; they generally lack integrated features such as persistent memory, real-time web access, voice interaction, and tool use that define the current consumer chatbot experience; and they require hardware and technical proficiency that most ordinary consumers do not have. Nor does local deployment eliminate all privacy risks. As I have argued elsewhere, even when a model runs entirely on local infrastructure, risks can persist: See Theodore Christakis, "DeepSeek and the China Data Question", IAPP (2025). The point here is not that local deployment is a superior alternative to cloud-based consumer chatbots. It is that its existence, and its growing adoption among privacy-conscious users, confirms that the trust deficit documented in this study has real competitive consequences.

What Do Users Actually Expect?

This study documents what providers reserve the right to do. But what about the other side of the gap? What confidentiality do users actually expect when they type into a chatbot? The honest answer is that the empirical base remains thin. No large-scale, nationally representative study has yet directly measured, across demographics and jurisdictions, what ordinary chatbot users believe happens to their conversations after they press "send".

That absence is itself a finding, and a call to action: rigorous, cross-jurisdictional survey research on user privacy expectations in the chatbot context is overdue and would substantially strengthen the normative foundations of the recommendations this study advances.

What evidence does exist, however, points consistently in one direction. A 2024 Consumer Reports survey of over 2,000 US adults found that nearly half (45%) of Americans believe companies operating chatbots should never store health-related information at all; only under 29% accepted use for training; and only 5% accepted sale or sharing for purposes affecting the consumer, such as targeted advertising.[18] A 2024 Deloitte survey of nearly 4,000 US consumers found that 62% of generative AI users were willing to discuss personal medical topics with a chatbot, and that a majority expressed interest in using chatbots for mental health support, relationship advice, and financial guidance, yet the same respondents identified data privacy and security as the primary condition for trusting the technology.[19] And a UK study of LLM chatbot users found that 76% of participants lacked a basic understanding of the privacy risks involved in their interactions, with many holding "flawed or outdated conceptions" of what happens to the data they share.[20]

The pattern these studies sketch is consistent with the structural gap this paper documents, and it suggests that the gap is not narrowing. Users are confiding more, in more sensitive domains, while their understanding of what happens to those disclosures remains poor and their trust in providers' data stewardship is, if anything, declining. The appropriate response is not fatalism. It is not enough to note that users over-disclose and shrug. The combination of growing intimacy and persistent ignorance is precisely the condition that demands better transparency, better defaults, and better design. It also demands investment, by researchers, regulators, and providers themselves, in sustained public awareness efforts that help users understand, in concrete and accessible terms, what chatbot privacy actually looks like today, so that they can make informed choices rather than drifting into acceptance of a status quo they have never been given the tools to evaluate.

Is this Just the Search Engine Debate All Over Again?

A natural objection is that consumer chatbot privacy raises nothing fundamentally new. Search engines, too, collect user queries, retain them, use them for product improvement

[18] Consumer Reports, "AI Chatbots: Two Consumer Reports Nationally Representative Surveys" (January 2024).
[19] Deloitte, "Earning Trust as Gen AI Takes Hold: 2024 Connected Consumer Survey" (September 2024). Younger cohorts (Gen Z and millennials) were roughly twice as likely as older generations to express interest in chatbot "friendship" (44% vs. 23%) and mental health counselling (44% vs. 24%).
[20] Malki et al., "Hoovered Up as a Data Point': Exploring Privacy Behaviours, Awareness, and Concerns Among UK Users of LLM-Based Conversational Agents", Proceedings on Privacy Enhancing Technologies (PoPETs) (2025). See also Lee et al. (2025), "Understanding Users' Privacy Reasoning and Behaviors During Chatbot Use", arXiv:2601.18125, which builds on the same body of work.

and advertising, and have been the subject of two decades of regulatory and judicial attention in Europe and beyond. If the privacy framework developed for search already addresses these concerns, the argument runs, then chatbots simply inherit the same rules and the same solutions.[21]

The objection is partly right and partly wrong, and getting the distinction right matters for both analysis and policy. It is right that certain structural features are shared. Both search engines and chatbots process user inputs on remote servers; both retain data for service improvement and safety; both may use interaction data for advertising; and both are subject, when operating in Europe, to the GDPR's full apparatus of lawful bases, purpose limitation, transparency, data minimisation, and data subject rights. The legal infrastructure built around search, including the CJEU's landmark *Google Spain* decision recognising a right to de-referencing, the ePrivacy Directive's consent requirements for tracking and behavioural advertising, and the EDPB's guidelines on search engine obligations, provides a body of doctrine that is relevant and, in places, directly applicable to chatbot providers.[22]

But the analogy breaks down at the point that matters most: the nature, depth, and intimacy of what the user discloses. A search query is typically short, fragmented, and instrumental. A user who types "symptoms headache nausea" into a search engine is disclosing a concern, but not narrating it. A user who opens a chatbot and writes "I've been having headaches and nausea for three weeks, I'm terrified it might be something serious, I haven't told my wife because she's already stressed about her mother's diagnosis, and I can't afford to see a specialist until next month" is doing something qualitatively different. The chatbot interaction is not a query; it is a confession. It contains not just a topic but a narrative, an emotional state, a family context, a financial constraint, and a medical fear, woven together in a form that the user would not share with a search bar and might not share with anyone else.

This difference is not a matter of degree; it is a difference in kind, and it has consequences across every dimension this study examines. At least five structural divergences separate the chatbot privacy problem from the search engine debate.

First, disclosure depth and narrative coherence. Search queries are atomistic: each query is typically self-contained, short, and disconnected from the next. Chatbot conversations are sustained, multi-turn dialogues in which the user builds a detailed, coherent narrative over minutes or hours, often across multiple sessions. As stated earlier, the empirical literature confirms that users disclose personally identifiable information, health data, relationship details, and professional secrets in chatbot conversations at a depth and frequency that has no equivalent in search query logs. A leaked or reviewed search query reveals a topic; a leaked or reviewed chatbot conversation reveals a life.

Second, the interface dynamic. Search engines present themselves as tools: the user types a question, receives a list of links, and leaves. The interaction is transactional, and

[21] The analogy to search has been drawn in several contexts. Google's own Gemini Apps Privacy Hub borrows directly from search-era warning language. The CJEU's foundational decision in *Google Spain SL v. AEPD* (Case C-131/12, 13 May 2014) established that search engine operators are data controllers under EU law and recognised a right to de-referencing. The solutions developed in that line of case law, however, were calibrated to the specific characteristics of search: short queries, indexing of third-party content, and the linking function. They do not map directly onto the chatbot context, where the provider is both the recipient and the processor of the user's own intimate disclosures.

[22] On the retention and reuse of search queries, see CJEU, Case C-131/12 *Google Spain* (2014), paras. 80-88, and Case C-136/17 *GC v. CNIL* (2019). On the right to erasure under Article 17 GDPR as applied to search engines, see EDPB Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (7 July 2020).

the interface does not invite further disclosure. Chatbot interfaces, by contrast, are designed to feel like conversations with a knowledgeable, empathetic interlocutor. They use natural language, maintain conversational continuity, remember prior exchanges, and adopt a counsel-like tone. This design choice is not incidental; it is the core product proposition. But it systematically encourages a level of self-disclosure that the search bar never invited, and it creates user expectations of confidentiality that the search engine never triggered. The privacy problem is therefore not just about what the provider does with the data, but about what the interface led the user to believe about the relationship.

Third, retention and reuse architecture. When a search engine retains a query, it retains a string of keywords. When a chatbot retains a conversation, it retains a rich, structured record that may include the user's emotional state, decision-making process, confidential facts, and responses to follow-up questions. This record can be used not only for training but also for personalisation ("memory"), human review, safety classification, and, as of 2026, advertising. The reuse surface is categorically larger than for a search query, and so is the potential harm from breach, compelled disclosure, or insider access. The solutions developed for search, such as query anonymisation after a defined period, do not address the distinctive risks created by retaining multi-turn, narrative-rich conversations for years.

Fourth, the training feedback loop. Search engines use query data primarily for ranking, autocomplete, and advertising. They do not, in general, use individual queries as training data for a model that will then reproduce or recombine the substance of those queries in outputs served to other users. Chatbot providers do exactly this: consumer conversations are used, by default, to train future model iterations. This creates a novel risk with no search-engine equivalent: the possibility that a user's original ideas, unpublished research, or confidential strategy could be memorised by the model and, under certain conditions, reproduced in responses to other users. The memorisation and regurgitation literature, including recent findings that production models can reproduce thousands of words from their training data, confirms that this is not a theoretical concern. Search engines never raised this problem because they index and link to existing content; they do not learn from user inputs and generate new outputs based on them.

Fifth, the human review dimension. Both search engines and chatbots may expose user data to human employees for quality, safety, or support purposes. But the content of what is exposed is radically different. A human reviewer who sees a search query sees a topic. A human reviewer who sees a chatbot conversation sees a detailed, often emotionally charged narrative that may include medical conditions, family conflicts, legal strategies, and financial distress, frequently combined in a single exchange. The privacy harm from human review is therefore of a different order, and the governance safeguards, including this study's proposal for a Sealed Mode, respond to a problem that search engines did not create at this scale.

None of this means that the search engine experience is irrelevant. On the contrary, several of the regulatory tools developed for search provide a useful starting point. The principle that users should be able to request deletion of personal data (the right to erasure under Article 17 GDPR, shaped by *Google Spain*) applies to chatbot data as well, and providers should ensure that deletion requests are honoured promptly and comprehensively. The consent requirements for behavioural advertising developed through the ePrivacy Directive and the CJEU's case law are directly relevant to any chatbot

that introduces ad personalisation.[23] And the transparency obligations that regulators have enforced against search engines, requiring clear disclosure of data retention, third-party sharing, and profiling, are at least as urgent in the chatbot context, where the stakes of opacity are higher.

But transposing search-era solutions without adaptation would be a category error. The privacy boundary for a technology that invites intimate, sustained, narrative self-disclosure is not the same as the privacy boundary for a technology that processes fragmented keyword queries. The solutions must be proportionate to the disclosure: deeper intimacy requires stronger constraints. That is why this study's recommendations go beyond what search-era regulation required, proposing not only transparency and consent mechanisms but also Sealed Mode, constraint-based privacy through cryptographic hardening, and governance rules designed for a product whose interface invites the user to confide, not merely to search.

A final observation: the boundary between search and chat is itself eroding. OpenAI now offers ChatGPT Search; Google has integrated Gemini into its search experience through AI Overviews and AI Mode; and analysts project that conversational AI could capture a significant share of traditional search volume within the next few years.[24] As these products converge, the privacy frameworks will need to converge as well, but in the direction of the higher standard, not the lower one. A system that combines the reach of a search engine with the intimacy of a chatbot should be governed by rules calibrated to the more sensitive of the two interactions, not by rules designed for the less sensitive one.

The 10 Recommendations: Scope and Ambition

The ten recommendations advanced in this study reflect both of these insights: that chatbot privacy builds on, but cannot be reduced to, the frameworks developed for earlier technologies, and that the trust problem is too urgent to wait for a comprehensive regulatory settlement.

They are deliberately framed as a starter set rather than a closed canon: workable, contestable, and open to refinement. Some may require adaptation to a provider's architecture, safety model, or market structure; others may prove easier to implement than expected once treated as product requirements rather than policy language. Many additional recommendations could and should be proposed, including by engineers, privacy scholars, consumer psychologists, and regulators. One ambition of this series is precisely to initiate a broader academic and policy debate about a foundational and still under-studied question: how to design consumer chatbot privacy boundaries that users can understand and rely upon, without stripping the systems of the benefits that made them indispensable.

The direction of travel is intentionally modest. It does not demand an unrealistic ideal of secrecy or a world without safety monitoring. It asks instead for a set of minimal, achievable commitments: clearer boundary disclosures that match the intimacy of the

[23] On the ePrivacy Directive's application to cookies and behavioural tracking in the search context, see Directive 2002/58/EC, Article 5(3), and CJEU, Case C-673/17 *Planet49* (1 October 2019). The CJEU confirmed that placing cookies for advertising purposes requires prior, active consent. On the advertising-related obligations applicable to search, see also CJEU, Case C-252/21 *Meta Platforms v. Bundeskartellamt* (4 July 2023), establishing stringent consent and necessity requirements for behavioural advertising by dominant platforms.

[24] On the convergence of search and conversational AI, see TechPolicy.Press, "The Privacy Challenges of Emerging Personalized AI Services" (May 2025), noting that AI services will set off "a race among providers for massive amounts of detailed user information, much of which is highly sensitive".

interface; more meaningful controls that do not force users to trade basic functionality for privacy; stronger internal constraints where disclosure risks are highest; and governance choices that prevent commercialisation or operational convenience from eroding confidentiality by stealth. In short, the claim is not that providers must choose between utility and privacy. The claim is that the long-term value of consumer chatbots depends on making their confidentiality boundary predictable, defensible, and worthy of the trust they increasingly solicit.

Outline

The present Part I addresses four questions that define the controller's internal boundary of use:

1) **Training and improvement**. To what extent may the provider treat a consumer interaction as material for training and product development, and what does an opt-out actually accomplish? What is excluded from the opt-out, and what remains visible to the provider even after the user exercises every available control?
2) **Human review**. Under what conditions may humans (employees, contractors, or third-party annotators) access consumer interactions for safety, quality, or support purposes, and how clearly is this disclosed to users in the interface rather than buried in policy text?
3) **Commercialization**. Are interactions used for advertising-adjacent purposes, profiling, or other forms of commercialization, and do providers clearly distinguish "improving the service" from "monetizing the conversation"?
4) **Operational sharing and ecosystem spillover**. Even where providers promise not to sell user data, how widely can chat content and derived signals circulate through the provider's operational stack (vendors, contractors, internal tooling, affiliates), and to what extent is the consumer chatbot experience architecturally "standalone" as opposed to integrated by default into broader service ecosystems?

Because consumer plans, defaults, and controls evolve quickly, the series flags significant plan and policy changes whenever they bear directly on privacy expectations,[25] and it treats ambiguity as a substantive finding rather than an accident. When a provider's documentation does not allow a user to determine, with reasonable precision, whether a given interaction may be reused, reviewed, retained, or commercially leveraged, that opacity itself becomes a privacy risk because it predictably invites over-disclosure. The sections that follow therefore aim to translate dense, fragmented documentation into a set of clear boundary lines, so that readers can understand, in concrete terms, what a consumer chatbot interaction is and is not today: a helpful tool, often, but not yet a relationship protected by anything like professional secrecy.

[25] For example, Anthropic switched from opt-in to opt-out for training on consumer chat data in 2025; OpenAI introduced the "Temporary Chat" feature; Google expanded Gemini training consent to opted-in teens. See King et al. (fn. 3) for a detailed account of these shifts.

I.
Your Words, Their Data ?
Training Models With Consumer Chatbot Conversations

1. Introduction: What Is Being Measured, and Why It Matters

This section examines the most consequential internal boundary in consumer chatbot privacy: whether, and how, the provider may treat a user's interaction as material for model training and product development. The question is not merely whether a conversation is stored, but whether it can become a learning signal that shapes future systems. In consumer settings, "improvement" typically covers several distinct practices: (i) using conversations as training data for future model iterations, (ii) using them for post-training processes such as reinforcement learning and alignment, (iii) using them to build or refine safety classifiers and abuse-detection systems, and (iv) using them for broader product analytics, debugging, and feature development. Providers routinely bundle these practices under a single label ("help improve"), even though they carry different privacy implications and are subject to different legal bases under frameworks such as the GDPR.

The central research task is therefore practical: for each provider, what is the default status for training on consumer chats, how can a user opt out, and what does that choice actually accomplish? Equally important is the design question that determines whether the choice is meaningful: does opting out impose a penalty on functionality, such as losing conversation history, personalization, or cross-device continuity? Where opting out comes at the cost of losing core usability, the voluntariness of consent is materially diminished. As the Stanford HAI study observed, the incorporation of user data for training by default is a practice that is essentially unique to consumer products; enterprise users are typically excluded from model training unless they explicitly opt in.[26]

2. Comparative Table

Table 1 maps the three variables that together determine whether a consumer chatbot interaction becomes training material: (i) the provider's default setting for training use, and any differences between free and paid tiers; (ii) the path a user must follow to opt out; and (iii) the penalty, if any, that opting out imposes on ordinary functionality such as conversation history, personalization, or cross-device continuity. Where a provider also offers an ephemeral or "private" chat mode as a separate safeguard, this is noted. The table is based on each provider's current consumer-facing documentation and in-product disclosures as of February 2026.

[26] King et al. observe that "[t]he incorporation of user data for model training by default is a practice unique to consumer products. Based on our analysis, enterprise users' data is typically excluded from model training." King et al., "User Privacy…", at 1472. See also King's congressional testimony: "Safeguarding Data Privacy and Well-Being for AI Chatbot Users" (18 Nov. 2025).

Table 1: Default Training Status, Opt-Out Paths, and Functionality Penalties Across Five Consumer Chatbot Services (as of 23 February 2026)

Provider	Default Status (Training Use)	Opt-Out Path	Penalty on Functionality
ChatGPT (OpenAI)	ON by default for all consumer tiers (Free, Go, Plus, Pro). Temporary Chat feature is not used for training. Feedback (thumbs up/down) can override the opt-out for the associated conversation.	Settings > Data Controls > toggle off "Improve the model for everyone"; or privacy portal ("do not train on my content"); or use Temporary Chat feature in the top right bubble in every chat window.	No loss of chat history when training is disabled; chats remain visible. No stated restrictions. Temporary Chat does not appear in history, not used for model training, and does not create Memory.
Gemini (Google)	Generally ON when Gemini Apps Activity ("Keep Activity") is ON. Temporary chats are not used for training. When Activity is OFF and no feedback is submitted, future chats are not used to improve Google's AI models. Conversations stored for up to 72 hours even when Activity is off.	Turn off Gemini Apps Activity in Google's activity controls; avoid submitting feedback for strictest posture; use temporary chats.	Turning off Keep Activity prevents chats from being saved in Gemini Apps Activity (other than short-term storage). Outside the EEA and UK it also disables features relying on saved chats (e.g., personalization).
Claude (Anthropic)	Since September 2025: ON by default for consumer plans (Free, Pro, Max) via the "Help Improve Claude" toggle. Before this date, Anthropic did not train on consumer chats unless feedback was submitted. Incognito chats are never used for training. Opting in extends retention to 5 years; opting out preserves 30-day retention.	Settings > Privacy > toggle off "Help Improve Claude". Incognito mode for individual sessions. The toggle was introduced within a mandatory policy-acceptance flow (deadline: 28 September 2025). For new users, the training choice is also surfaced directly during the onboarding flow.	No explicit functionality penalty stated for disabling Model Improvement. Incognito chats are not retained in history. However, the retention differential (5 years vs. 30 days) effectively makes the training choice also a retention choice, although user-initiated deletion of conversations overrides the extended retention period.
Grok (xAI/X)	ON by default. Grok on X uses X's public data and user interactions for training unless disabled in privacy settings. Standalone grok.com has separate controls. Private Chat ("Temporary Chat") is not used for training. Feedback (thumbs up/down) can override the opt-out, as with OpenAI.	On X: Settings > Privacy & Safety > Grok & Third-party Collaborators > uncheck training. On grok.com: privacy settings. Private Chat as session-level alternative.	Opting out does not remove Grok access; user can still interact. Private Chat does not show conversation history. Recent X ToS revision (Jan. 2026) expands "Content" to include prompts/outputs with broad reuse license.
DeepSeek Chat (DeepSeek)	Policy states conversations may be used to train and improve models; default not clearly operationalized in public UI disclosures.	Policy states users can opt out of AI training; specific UI path and conditions are not clearly described.	Not clearly disclosed: consumer-visible trade-offs (history, personalization, continuity) are not described with sufficient precision.

3. Analysis: What 'Training and Improvement' Actually Entails

Across providers, the phrase "improve the model" is doing heavy work. In consumer documentation it routinely collapses several distinct processes: continued training on conversations, post-training alignment, fine-tuning safety and policy models, and more general product-development work. A well-designed user control should specify, in plain language, which of these processes the opt-out affects. King et al. go further, arguing that the framing of the training opt-in itself can be manipulative: OpenAI presents it as "Improve the model for everyone," a formulation that invokes social guilt rather than explaining the concrete data use.[27]

OpenAI's Data Controls are unusually explicit on one key point: turning off training does not remove chats from history and imposes "no restrictions."[28] This makes the user choice closer to a genuine preference than a forced trade. However, this clarity is undercut by a significant caveat that is easy to miss: even after opting out, if the user clicks a thumbs-up or thumbs-down button on any response, the *entire conversation* associated with that feedback may be used for training.[29] For a user who has carefully opted out but routinely provides feedback, the opt-out is functionally partial.

Google's design creates a major practical inconvenience for users who want to opt out of training or "improvement" uses while still being able to rely on ordinary continuity. In practice, many people use chatbots as a working space: they return to a thread to refine a work project, revisit a health-related conversation, or simply pick up where they left off. Yet Google places heavy weight on the "Keep Activity" setting: turning it off generally stops chats from being saved in Gemini Apps Activity (beyond short-term retention), which means the user loses the ability to easily continue, build on, and rework prior conversations inside the service. The result is a structural dilemma: if a user wants the normal convenience of chat history and continuity, they may feel pushed to keep activity on even if that also entails broader downstream uses of their chats for model improvement. This is not just a "setting choice"; it is a design trade-off that effectively ties everyday usability to acceptance of wider processing.[30]

Anthropic's approach deserves extended attention because it illustrates how rapidly the privacy landscape can shift. Until September 2025, Anthropic did not train on consumer chat data at all, a policy that had become a significant selling point for privacy-conscious users. The September 2025 update reversed this default: all consumer plans (Free, Pro,

---

[27] OpenAI frames the training opt-in in social terms: "Improve the model for everyone: Allow your content to be used to train our models, which makes ChatGPT better for you and everyone who uses it." King et al. identify this as an example of "guiltshaming": "The framing of this appeal in social terms rather than individual (e.g., 'improve the model', or 'improve the model for you') is designed to invoke users' guilt… and attempts to persuade users by aligning the company's interests with the public good." King et al., at 1474.

[28] OpenAI Help Center, "How Your Data Is Used to Improve Model Performance." See also OpenAI, "Data Controls FAQ" (confirming that turning off model training preserves chat history: "Your conversations will still appear in your chat history but won't be used to train ChatGPT. [Y]ou can change this setting anytime. There are no restrictions"). Temporary Chats are deleted from OpenAI's systems within 30 days.

[29] A critical and easily overlooked caveat: OpenAI states that "[e]ven if you have opted out of training, you can still choose to provide feedback to us about your interactions with our products (for instance, by selecting thumbs up or thumbs down on a model response). If you choose to provide feedback, *the entire conversation associated with that feedback may be used to train our models*". This means a single thumbs-up or thumbs-down click can override the account-level opt-out for that conversation. See OpenAI, "How Your Data Is Used to Improve Model Performance" (emphasis added).

[30] Google, "Gemini Apps Privacy Hub." Temporary chats are not used for training. When Gemini Apps Activity is turned off and no feedback is submitted, future chats are not used to improve Google's AI models. Conversations are stored for up to 72 hours even when Activity is off for safety and abuse prevention.

Max) are now subject to training use unless the user actively disables the "Help Improve Claude" toggle.[31] For existing users, the toggle was presented within a mandatory policy-acceptance flow; for new users, the training choice is also surfaced during onboarding, making it one of the more visible opt-out paths among the providers studied. Anthropic's consumer documentation does provide a distinctive and unusually precise account of what the opt-out does and does not do. Turning off Model Improvement stops the use of new chats for future training, but may not affect training processes already underway or models already trained. Anthropic also highlights an important carve-out: conversations flagged by safety classifiers may still be used to improve trust-and-safety models, enforce policies, or advance safety research.[32] This is an honesty point: many users interpret "opt-out" as an absolute shield. It rarely is.

A further dimension of Anthropic's approach concerns retention. Users who opt in to Model Improvement consent to backend data retention of up to five years; those who opt out retain the previous 30-day backend retention window. These periods govern how long data may persist in Anthropic's back-end systems (for example, de-identified training pipelines), and do not affect users' ability to access their conversations in-product. Importantly, if a user deletes a conversation or their account, that deletion overrides the five-year retention period. Anthropic has published a public FAQ explaining the rationale for these retention windows.[33] Nevertheless, the opt-in/opt-out decision is not solely about training: it is also a retention decision, and the link between the two choices may not be immediately apparent to all users.[34]

Users who opt in to Model Improvement consent to data retention of up to five years; those who opt out retain the previous 30-day retention window. This means the opt-in/opt-out decision is not solely about training: it is also, and perhaps more consequentially, a retention decision.

Retention periods must be justified and proportionate under the GDPR's storage-limitation principle (Art. 5(1)(e)); indefinite or very long retention of chat data raises distinct data-protection risks, regardless of whether the data is used for training.

For Grok, the opt-out is presented as a straightforward privacy setting on X: disabling training and fine-tuning using public data and Grok interactions.[35] The X help text

---

[31] The September 2025 policy shift is documented in Anthropic's updated Consumer Terms and Privacy Policy (effective 28 September 2025). See also "How do I change my model improvement privacy settings?". Prior to this date, Anthropic did not use consumer chat data for training unless users submitted explicit feedback. The update introduced a "Help Improve Claude" toggle, presented within a mandatory policy-acceptance flow. Users who enabled the toggle (or accepted the new terms without disabling it) consented to training use going forward; those who opted out retained the prior 30-day retention policy. See MacRumors, "Anthropic Will Now Train Claude on Your Chats" (28 Aug. 2025); AMST Legal, "Anthropic's Claude AI Updates" (25 Sept. 2025). King et al. note that after this change, "all six of the developers now train on their users' chat data by default."

[32] Anthropic Privacy Center, "How Do I Change My Model Improvement Privacy Settings?." See also Anthropic Help Center, "Who Can View My Conversations?" (confirming the safety-classifier exception: "If our safety classifiers flag your conversations, they may still be used to improve our internal trust and safety models, detect harmful content, enforce our policies, or advance our safety research."). Incognito chats are not used to improve Claude, even if Model Improvement is enabled.

[33] See https://www.anthropic.com/news/updates-to-our-consumer-terms.

[34] Anthropic states that opting in to Model Improvement extends data retention to up to five years, while opting out preserves the prior 30-day retention period. This difference is significant: it means the opt-in/opt-out decision is not only about training but also about how long Anthropic stores the conversation data. See gHacks, "Anthropic's New Policies" (29 Aug. 2025); Tom's Guide, "Your Claude Chats Are Being Used to Train AI" (1 Oct. 2025). Note also that reopening an older conversation after accepting the new terms converts it into a "resumed" session subject to the current training setting.

[35] X Help Center, "About Grok." The opt-out path is: Settings and Privacy > Privacy and Safety > Grok & Third-party Collaborators > uncheck "Allow your public data as well as your interactions, inputs, and results with Grok and xAI to

confirms that users can still use Grok if they opt out. However, two complications deserve mention. First, as with OpenAI, voluntary feedback (thumbs up/down) can re-enable training use for the associated conversation, even after opting out.[36] Second, a recent revision to X's Terms of Service (effective 15 January 2026) expands the definition of user "Content" to expressly include prompts, outputs, and information obtained or created through the Services, granting X a worldwide, royalty-free, sublicensable license to use this content "for any purpose" — including training AI models. The scope of this license would appear to undercut the granular opt-out control offered by the privacy setting.[37] Private Chat ("Temporary Chat" on grok.com) avoids training use but also removes conversation history.

DeepSeek's policy text describes training and an opt-out right, but without a similarly operational description of the setting or its scope, users cannot accurately predict the real boundary.[38]

4. Opt-Out Semantics: What the User Should Assume

Three clarifications are essential for consumer comprehension. First, "not used for training" is typically forward-looking: it stops (or aims to stop) the use of new conversations in future training runs, but cannot reliably retract the influence of data already incorporated into a model's weights. Second, most providers reserve the right to use conversations for security and safety purposes even when model improvement is disabled, including to detect abuse, investigate incidents, and train safety classifiers. These are not merely legal boilerplate; they materially affect what "privacy" means in practice, especially for users whose content is likely to trigger safety review. Third, and this point has received insufficient attention, at least two providers (OpenAI and xAI) allow a single feedback action (thumbs up or thumbs down) to override the account-level opt-out for the entire associated conversation. The practical result is that a conscientious user who has opted out of training but who uses the feedback buttons is unknowingly re-consenting on a conversation-by-conversation basis.

5. Can Original Ideas Be 'Regurgitated' to Others? What the Chatbots Themselves Admit

A recurring academic and professional concern is whether creative or original ideas shared in a chatbot conversation can reappear in responses served to other users. The

be used for training and fine-tuning." See also xAI Consumer FAQs (Private Chat not used for model training; 30-day deletion window for deleted conversations and Private Chat).

[36] Like OpenAI, xAI reserves a feedback-based override: "Please also note that you can voluntarily submit feedback on a conversation even if you opt out of model training using, for example, the thumbs up and thumbs down buttons. If you provide such feedback, X and xAI may use that conversation and any such conversation related data to train and fine-tune Grok." See X Help Center, "About Grok". This caveat appears only on the help page, not on the opt-out interface itself.

[37] Grok's privacy picture is further complicated by the distinction between Grok accessed through the X platform (governed by X's Terms of Service and privacy settings) and the standalone Grok app (grok.com, governed by xAI's Consumer Terms). The recent X Terms of Service revision (effective 15 January 2026) expands the definition of user "Content" to include "inputs, prompts, outputs, and information obtained or created through the Services," and grants X a worldwide, royalty-free, sublicensable license to use this content "for any purpose", including training AI models. Users are responsible for Content and X cautions users to "only provide, create, or generate what they are comfortable sharing". See CryptoSlate, "X Claims the Right to Share Your Private AI Chats" (17 Dec. 2025).

[38] DeepSeek Chat, Privacy Policy. The policy states conversations may be used to train and improve models and that users can opt out; however, the specific UI path, operational conditions, and any functionality trade-offs remain opaque in public documentation.

short answer is that no provider offers a strict non-reproduction guarantee for consumer chats. Instead, they rely on a combination of (i) minimization and filtering before training, (ii) privacy-preserving post-training techniques, and (iii) policies against deliberate disclosure of personal data. Even with those measures, the research literature demonstrates that large language models can memorize and reproduce portions of their training data under certain conditions, including via targeted extraction attacks.[39] This means that if a provider trains on consumer conversations, the possibility of future reproduction cannot be treated as purely hypothetical. The magnitude of this risk, however, varies significantly depending on how a given provider trains on consumer data and what mitigations it applies in the training process - including deduplication, filtering, differential privacy techniques, and output-level safeguards against verbatim reproduction.

The scale of this risk was underscored by a study published in February 2026 by researchers at Stanford and Yale, who demonstrated that leading closed models can be strategically prompted to reproduce thousands of words from bestselling novels.[40] By asking models to complete sentences from copyrighted books, researchers found that Gemini 2.5 regurgitated 76.8 per cent of *Harry Potter and the Philosopher's Stone* with high accuracy, while Grok 3 reproduced 70.3 per cent. Researchers were also able to extract nearly the entirety of a novel from Anthropic's Claude 3.7 Sonnet by jailbreaking the model. These results are significant because closed models, which tend to have stronger safeguards against generating unwanted content, were previously assumed to be less prone to large-scale memorization than open-weight models such as Meta's Llama. As one of the researchers observed, it was "a surprise that they could memorise entire texts" despite guardrails. While these findings relate to copyrighted literary works rather than to user chat data,[41] the underlying mechanism is the same: if models trained on consumer conversations memorize inputs with comparable fidelity, and if a given provider's training pipeline does not apply sufficient mitigations to prevent it, the privacy implications for users who share original ideas, unpublished research, or confidential strategies could be severe.[42] The risk is accordingly not uniform across providers, and

[39] Nicholas Carlini et al., "Extracting Training Data from Large Language Models" arXiv:2012.07805 (2020); Milad Nasr et al., "Scalable Extraction of Training Data from (Production) Language Models" arXiv:2311.17035 (2023). This study demonstrated extraction of memorized training data from aligned, deployed models (including ChatGPT), showing that production safety measures do not eliminate the risk.

[40] A. Feder Cooper et al., Stanford and Yale Universities (2026), reported in Cristina Criddle, "AI's 'Memorisation' Problem: The Novels It Can't Forget", *Financial Times*, 23 February 2026, . The study builds on earlier work on open-weight models: see Chang et al., "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4" (2023), which found extensive memorization of popular fiction in open models. The FT article also reminds the famous decision of a German court that ruled in November 2025 that OpenAI had infringed copyright because its model had memorized song lyrics (GEMA v. OpenAI), considered a landmark ruling in the EU.

[41] Feder Cooper et al. focus primarily on extraction of in-copyright books from production LLMs via a two-phase procedure that often begins by prompting continuation from a short known prefix and then iteratively continuing. This is best understood as a form of targeted extraction against memorized copyrighted text where the attacker has a public 'hook' (the prefix). For private user chat content, an external attacker typically lacks such ground-truth hooks, which can make targeted extraction of a specific individual's conversation structurally more difficult. That said, the paper remains significant for confidentiality analysis because it demonstrates that extractable memorization can persist in production systems even with guardrails; and related work has studied extractable memorization more generally, including settings where the adversary need not know the underlying training dataset in advance.

[42] Academic extraction results should be distinguished from ordinary consumer risk. The production-LLM extraction procedure reported by Feder Cooper et al. can involve large numbers of attempts (including Best-of-N techniques and repeated continuation prompts), making successful extraction potentially costly and expertise-intensive in practice, and the study's empirical focus is copyrighted books rather than the disclosure of personal user chat data. Publicly documented cases of harmful regurgitation of identifiable private chat data from mainstream consumer chatbots appear limited, and major providers explicitly design and train safeguards intended to prevent disclosure of personal

users would benefit from transparency about the specific safeguards each provider applies.

Two distinctions matter. First is immediacy: a model does not typically "learn" a new user's idea and instantly share it with the next user, because training is performed offline in periodic runs. However, immediate leakage can occur through other mechanisms (for example, when content is routed for human review, shared internally, or included in a dataset used by evaluators). Second is indirect regurgitation: after future training, the model may become more likely to reproduce distinctive passages that were included in training data, especially if the content was duplicated, unusually specific, or repeated across contexts. Deduplication of training data significantly reduces but does not eliminate this risk.[43]

To illustrate how these risks are communicated in practice, the author conducted a simple **test** on 23 February 2026, posing the same question to each of the five chatbots examined in this article: **what is the risk that unpublished academic ideas discussed in a consumer chat session could leak to other users or be reused in future responses?** The full responses are reproduced in the Appendix[44].

The results are revealing, and more unsettling than the reassuring opening lines suggest. All five chatbots begin by characterising the risk as "very low", "extremely low", or "negligible". But none stops there. As the responses develop, each chatbot concedes, in its own way, that the risk is not zero and that the primary vector is the training pipeline: conversations used to improve future models can, in principle, influence what those models say to other people. What is striking is how explicit several of the chatbots are about the mechanism. ChatGPT warns that if a user shares "a novel framing, taxonomy, catchy terminology, or a particularly good structure, there is a small chance a future answer to someone else on a related topic echoes some of that framing without knowing its source", and adds that the risk is "higher" when the user shares "signature elements: a new named concept, an original framework, a detailed argument map, a unique empirical strategy". Claude acknowledges "a small theoretical possibility that distinctive formulations or novel conceptual frameworks could influence model outputs in subtle ways". Grok describes a "moderate-but-still-small" risk of "generalized influence on the model's knowledge in that narrow domain". DeepSeek is the most candid: it identifies the training pipeline as "the primary mechanism through which your ideas could theoretically influence future responses to other users", states that "the inherent nature of how AI models learn and operate creates a tangible possibility that your novel ideas could be incorporated into the model's knowledge base and reflected in responses to other users", and advises users to "adopt the mindset that anything you type could

data even when prompted. This does not eliminate the risk in principle, but it helps separate high-effort proof-of-concept attacks from the baseline threat model faced by most users.

[43] Katherine Lee et al., "Deduplicating Training Data Makes Language Models Better" arXiv:2107.06499 (2021). Demonstrates that deduplication significantly reduces memorization, but does not eliminate it.

[44] A methodological caveat is necessary. Large language models are known to hallucinate (see my previous articles on this topic), and their self-reported descriptions of their own data practices cannot be treated as reliable statements of fact. This section does not rely on the chatbot responses as a source of technical truth about provider behaviour; for that, the study relies on published documentation, policies, and interface disclosures examined in the preceding sections. The purpose of the exercise is different: it examines what users are told when they ask the product itself about privacy risk, and whether the pattern of reassurance followed by a battery of protective recommendations reveals a gap between the confidence of the initial answer and the caution of the practical advice. That pattern, which is consistent across all five chatbots, is the finding, and it holds regardless of whether any individual factual claim within the responses is accurate. Indeed, if a chatbot provides inaccurate information about its own data practices when directly asked by a user, that is itself a transparency concern.

potentially be read by others" and not to share an idea they "wouldn't be comfortable seeing on a public blog or in a competitor's grant proposal"!

What is absent from the responses is at least as instructive as what is present. None references the empirical memorization literature or the recent findings that production models can regurgitate near-verbatim passages from their training data. And the levels of self-knowledge varied strikingly: Claude provided specific, verifiable details about its own data practices (the date of the policy change, the retention periods, the settings path); Grok disclosed an actual data breach incident involving approximately 370,000 publicly indexed chat transcripts; DeepSeek cited a real-world example where "confidential project information fed into the system could be recalled and attributed to the user in a completely separate session".

Perhaps the most telling finding is this: every chatbot, having begun by reassuring the user that the risk is low, proceeded to recommend a series of protective measures whose collective weight sits in tension with that reassurance. These recommendations range from the procedural (opt out of training, use private or incognito chat modes, delete conversations after use, check privacy settings) to the substantive (anonymise and generalise, use placeholders instead of real data, share only "high-level direction" rather than specific contributions, keep distinctive vocabulary private, call a novel concept "Concept X" rather than revealing its name). Several go further still, advising users to consider moving sensitive work to local, offline models that never transmit data to a server, or to treat the chatbot as a "public-access tool". When the product itself advises users to withhold their best ideas, the gap between the confidentiality that users might reasonably expect and the confidentiality that actually exists is no longer a matter of inference. It is stated, in the chatbot's own words.

5. Conclusions, Lessons Learned, and Recommendations

First, default settings matter more than policy text. In all five services, the baseline consumer posture permits some form of improvement use unless the user actively changes settings or selects a special mode. Since Anthropic's September 2025 reversal, *every major provider* now trains on consumer chat data by default — a finding confirmed by King et al. across a broader six-provider sample.

Second, opt-out is rarely an absolute: even when model training is disabled, providers typically reserve safety and abuse-prevention uses, and at least two providers allow feedback actions to override the opt-out.

Third, the most consumer-protective designs are those that allow users to disable training without losing ordinary usability. OpenAI's disclosure that turning off training preserves chat history and carries 'no restrictions' is a strong benchmark for meaningful choice. Gemini illustrates, by contrast, a practical trade-off created by a single control. Google's "Keep Activity" setting both enables chat history (so users can return to and continue earlier threads) and, when enabled, permits the stored Gemini Apps Activity to be used for the purposes described in Google's Privacy Hub, including improving and developing services and training generative AI models. Users who would prefer to minimise such uses may therefore feel compelled to leave activity on simply to preserve continuity for ongoing work, health-related questions, or other important conversations. Anthropic's retention differential introduces a further distortion: if a user enables model training, Anthropic may retain data in de-identified training pipelines for up to five years,

whereas deleted data is removed from back-end systems within 30 days - though Anthropic specifies that user-initiated deletion of a conversation or account overrides the five-year period - making the training choice difficult to disentangle from the retention choice, in ways that may affect the specificity of consent under the GDPR.

Lessons Learned

Consumers routinely interpret "chat" as confidentiality. Providers should assume over-disclosure and design controls accordingly. The self-disclosure literature confirms that perceived anonymity in chatbot interactions encourages users to share information they would withhold in other contexts.

The real privacy boundary is the combination of policy + interface: an opt-out that costs the user core features, or that can be silently overridden by a feedback click, is not functionally equivalent to an opt-out with no penalty. The EDPB Taskforce Report's insistence that controllers may not transfer compliance responsibility to users through terms of service should inform how these controls are evaluated.[45]

Training on consumer chats increases the long-term risk that distinctive content can reappear indirectly, even if immediate leakage is unlikely. The memorization literature confirms extractable regurgitation is a real, measurable phenomenon that persists even in production-aligned models.[46]

Recommendations

Recommendation 1: Decouple history, retention, and training. A basic privacy-by-design improvement is to allow full conversation history and cross-device continuity while keeping training disabled, without requiring users to accept expanded downstream uses or extended retention. Where chat history or personalization requires storing conversations, that storage should not automatically imply permission to use the content for model training, and the retention period should not vary depending on the training choice.

Recommendation 2: Make opt-out semantics explicit and feedback-proof. Providers should offer a small number of intelligible, user-facing choices that clearly distinguish (i) model training and post-training improvement, (ii) product analytics and debugging, and (iii) safety and abuse-prevention uses. A single "help improve" toggle collapses materially different purposes and undermines informed choice.[47] In addition, providers should not allow a single feedback action (thumbs up/down) to override an account-level opt-out; or, at a minimum, should present a clear, unavoidable notice before treating feedback-linked conversations as eligible for training or improvement.

---

[45] EDPB, Report of the Work Undertaken by the ChatGPT Taskforce (23 May 2024), at para. 23-26 (on fairness: "a crucial aspect of fairness is that there should be no risk transfer," meaning controllers should not shift GDPR compliance responsibility to data subjects, for example by arguing that users are responsible for the personal data they input).

[46] Mindy Nunez Duffourc, Sara Gerke & Konrad Kollnig, "Privacy of Personal Data in the Generative AI Data Lifecycle," 13 *NYU J. Intell. Prop. & Ent. L.* 238 (2024), available at https://jipel.law.nyu.edu/privacy-of-personal-data-in-the-generative-ai-data-lifecycle/ (discussing the legal treatment of user inputs and data retention across the GenAI lifecycle under both GDPR and U.S. frameworks).

[47] This is not a call for 'hyper-granular' settings, but for a limited set of stable, comprehensible purpose categories so that users can make a meaningful choice without navigating a maze of controls.

Recommendation 3: Adopt regurgitation-aware safeguards with documented residual risk. If consumer conversations are used for training, providers should publish a clear, verifiable account of mitigation measures (deduplication, filtering/obfuscation, privacy-preserving post-training methods) and the residual risk.[48] For high-stakes user groups — academics, lawyers, doctors, journalists, whistleblowers — providers should consider offering, to the extent possible, a clearly labeled mode where user content is not used for training and is subject to strict, independently verifiable retention limits.

[48] See Google DeepMind, Gemini 2.5 Model Report, Section 5.6 ('Memorization and Privacy'), which reports quantitative evaluations of both 'discoverable memorization' and 'extractable memorization' (including a divergence-style attack). The report indicates that across outputs classified as memorization for the Gemini 2.X model family, the authors observed no personal information (noting results are bounded by detection thresholds), and that in cases where divergence was successful the rate of emitted text classified as memorization was roughly 0.2%, with examples described as often boilerplate code or web content. These results support the value of provider reporting that quantifies both mitigations and residual risk.

II. The Ghost in the Machine: Human Review of Consumer Chatbot Conversations

1. Introduction: what is being measured, and why interface warnings matter

This section examines a privacy boundary that users intuitively care about more than almost any other: whether a human being can read their conversation with a chatbot. Even when a provider promises that chats are "not used for training", the user's sense of confidentiality can be defeated if employees or contractors may access the interaction for quality, safety, or support. The core questions are therefore: under what conditions do providers reserve human access; what purposes are invoked (quality assurance, safety monitoring, incident investigation, abuse enforcement, customer support); and, crucially, how clearly is this communicated in the interface rather than buried in policy text.

1.1 Gemini's interface warning as a transparency benchmark

Among the five services, Google's Gemini offers the most striking example of direct, consumer-legible disclosure. In its Gemini Apps Privacy Hub, Google states (in bold):

> *"Please don't enter confidential information in your conversations or any data you wouldn't want a reviewer to see or Google to use to improve our services, including machine-learning technologies."*[49]

This sentence is unusually candid. It does not merely acknowledge that human review exists; it translates the implication into an instruction: if a user would not want a reviewer to see the content, they should not type it. The transparency benefit is obvious, but so is the normative implication. When a product that feels like a private conversation must carry such a warning, the default user expectation of confidentiality is likely miscalibrated.

There is, however, a tension that must be acknowledged. The EDPB's ChatGPT Taskforce has established that a "crucial aspect of fairness is that there should be no risk transfer" from controller to data subject, and specifically that controllers may not "transfer the [compliance] responsibility to data subjects, for example by placing a clause in the Terms and Conditions that data subjects are responsible for their chat inputs."[50] Applied here, Gemini's warning is commendable as transparency, but it cannot serve as a substitute for the provider's own obligation to protect user data. Telling users not to type sensitive information does not relieve Google of its duty to process whatever users do type in accordance with the GDPR's lawfulness, fairness, and purpose-limitation requirements. The warning is a floor, not a ceiling.

2. Comparative table

Table 2 maps four variables that together define the human-review boundary for each consumer chatbot: (i) whether the provider issues an explicit interface warning about human review, or confines disclosure to policy text; (ii) the stated purposes for which human review may occur; (iii) whether the provider offers any guarantee that reviewed

[49] Google, Gemini Apps Privacy Hub, https://support.google.com/gemini/answer/13594961. The warning appears in the same informational environment as the product; human review and retention details are set out under "Who has access to my chats" and "How long is reviewed data retained."
[50] EDPB, Report of the Work Undertaken by the ChatGPT Taskforce (23 May 2024), para. 23-26.

data will remain confidential and will not be used for secondary purposes; and (iv) the retention period applicable to human-reviewed data, where disclosed.

Table 2: Under what conditions may humans access consumer chats, and how clearly is this disclosed in the interface? (as of 23 February 2026)

Provider	Explicit UI warning?	Stated reason for review	Confidentiality guarantees for reviewed data?	Retention of reviewed data
ChatGPT (OpenAI)	Mainly policy/help-center; interface messaging varies by surface.	Improve systems; ensure content complies with policies and safety requirements; investigate misuse; support.	No blanket confidentiality guarantee; describes review for safety/responsible AI and service improvement. Staff and third-party contractors bound by confidentiality agreements.	No distinct retention period for a "human-reviewed subset" clearly disclosed; Temporary Chats retained up to ~30 days unless exceptions.
Gemini (Google)	Both: explicit bold warning in Gemini Apps Privacy Hub + policy detail.	Provide service; improve Gemini and other generative AI models; keep Gemini safe and secure (with human reviewers); process feedback.	Chats "disconnected" from Google Account before review; warns users not to enter confidential info; no absolute confidentiality guarantee.	Reviewed chats retained up to 3 years, not deleted when user deletes activity; some Live chats stored up to 72 hours even with activity off.
Claude (Anthropic)	Primarily help-center/privacy materials (limited default UI warning).	Safety & abuse prevention; investigate security incidents & misuse; improve Claude depending on settings/feedback. Classifier-based real-time monitoring with human escalation.	Data de-linked from user ID before review; no promise of absolute confidentiality; safety exceptions acknowledged. Extended retention for flagged content.	Deleted chats removed within 30 days; flagged content retained up to 2 years; safety classification scores up to 7 years. Feedback data retained up to 5 years.
Grok (xAI/X)	Help-center/policy; no Gemini-style default warning.	Authorized personnel may review to improve model performance, investigate security incidents and misuse, and comply with legal obligations.	No absolute confidentiality promise; review limited to "authorized personnel" for "specific business purposes."	No distinct retention period for human-reviewed subset; Private Chat deleted within ~30 days unless legal/compliance/safety reasons.
DeepSeek Chat	Policy-level references; no clear UI warning in primary sources reviewed.	General security & operational purposes; access limited to authorized personnel when necessary.	No explicit confidentiality guarantee beyond access-limitation language.	No clearly specified retention period for a human-reviewed subset in primary policy.

3. Analysis: what the disclosures reveal about real confidentiality boundaries

Three patterns emerge from the comparison.

First, every provider preserves human access pathways, but they describe them with sharply different levels of candor. Gemini's warning functions as a transparency benchmark. It is not merely a legal disclosure; it is behavioral guidance that appears in the same informational environment in which users decide what to type. Other services acknowledge review in help pages or privacy materials, but without an equally direct, prominent warning. The consequence is predictable: many users will assume a level of confidentiality that the product does not actually guarantee. Empirical research confirms the risk: studies of real-world chatbot conversations reveal that users routinely share personally identifiable information, health data, relationship details, and professional secrets in contexts where human review is at least theoretically possible.[51]

Second, providers tend to frame human review as a limited subset ("a limited number of authorized personnel," "some chats," "to help keep the service safe"), but they rarely specify when a particular user's chat enters that subset. This creates a fundamental asymmetry. The user experiences a stable, intimate interface, while the review boundary is probabilistic and opaque. In consumer protection terms, this is a classic scenario of material information being technically available but practically undisclosed.

Third, the safeguards that providers emphasize are largely procedural but also sometimes technical. Gemini and Claude stress de-linking or disconnecting reviewed chats from account identifiers and also use automated filters designed to remove personally identifying information. That reduces identification risk, but it does not eliminate confidentiality harm, because the conversation itself can contain names, unique facts, and narrative details that are identifying in context. OpenAI and xAI describe review for safety and improvement, but provide fewer consumer-visible specifics about how review datasets are handled, who can access them, and how secondary uses are constrained. A recent NordVPN analysis of ChatGPT's access architecture identified at least six categories of personnel or systems with potential access to consumer conversations, including authorized employees, third-party contractors, automated classifiers, enterprise administrators, infrastructure providers (Microsoft Azure), and law enforcement under valid legal process.[52] DeepSeek's disclosures remain the least operational, leaving users with limited ability to assess when humans can see their content and what happens afterwards.

3.1 Retention as the hidden multiplier of human access

Retention determines how long a reviewed conversation remains available for secondary access. Here, Gemini's disclosure is unusually stark: chats reviewed by human reviewers can be retained for up to three years and are not deleted when users delete their activity.[53]

---

[51] Niloofar Mireshghallah et al., "Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild" COLM 2024. The study analysed real-world ChatGPT conversations shared on the ShareGPT platform and found that a substantial proportion contained personally identifiable information, sensitive health data, and professional secrets. See also SafetyDetectives research (September 2025), which found that leaked Grok and ChatGPT conversations contained PII, emotional disclosures, and confidential business material.

[52] NordVPN, "Is ChatGPT private? A 2026 guide to your data privacy and security", 13 January 2026.

[53] Google, Gemini Apps Privacy Hub : "Conversations that have been reviewed by human reviewers (and related data like your language, device type, location info, or feedback) are not deleted when you delete your activity. Instead, they

This creates a situation in which the user believes they have exercised control over their data, but a copy of the reviewed conversation persists in a separate, disconnected dataset that the user cannot reach. Anthropic provides relatively granular retention windows, including a two-year retention period for inputs and outputs flagged by trust-and-safety classifiers, and a seven-year retention period for the associated classification scores. The practical implication is that the conversation content flagged in 2026 could persist in Anthropic's safety infrastructure until 2028, while the associated classifier scores (numerical outputs that indicate why a conversation was flagged, but not the conversation itself) could persist until 2033. The longer retention of classifier scores serves operational purposes such as monitoring whether the rate at which a particular classifier fires is changing over time, but it means that a record of the fact that a user's conversation triggered a safety flag can outlast the conversation itself by a significant margin. Anthropic has also published a detailed account of the classifier architecture that triggers these retention extensions, including real-time safety classifiers, hierarchical summarization for detecting account-level abuse patterns, and partnership-based threat intelligence.[54]

Where providers do not disclose a distinct retention schedule for human-reviewed subsets, users cannot realistically assess the duration of exposure, and the risk of over-sharing increases accordingly. The asymmetry is compounded by the fact that users do not know, and typically cannot discover, whether any particular conversation has been selected for review.

4. Sealed Mode: When the Conversation Must Not Travel

A recurring theme in the comparative analysis is a growing mismatch between the intimacy the interface invites and the reality that a subset of consumer chats may be retained, sampled, and reviewed by humans for safety, quality, or support. For many ordinary uses, that governance model is defensible. The hard cases are different: situations where the expected harm from disclosure is exceptionally high, and where users increasingly treat a consumer chatbot as a substitute for a protected relationship. The clearest example is health and mental health. Millions of people already use chatbots to describe symptoms, discuss medication, process grief, or navigate moments of acute emotional distress.[55] They do so not because the chatbot promises medical confidentiality, but because the interface *feels* confidential. The gap between the experienced and the actual level of protection is wide, and it is widening as chatbots become more fluent, more continuous, and more counsel-like in tone.

This is not a problem that warning labels can solve on their own. Telling users "do not share sensitive information" while the product design continues to encourage disclosure through fluency, continuity, and a personalised counsel-like tone is, as the EDPB's

---

are retained for up to three years." This means a user who deletes all Gemini activity cannot remove conversations that have already been selected for human review. The three-year retention applies to the disconnected, reviewed copy, not to the user's account-linked activity. See also Gizmodo, "Google's Gemini AI Keeps Your Conversations for Up to 3 Years", and Search Engine Journal, "Google Gemini Warning: Don't Share Confidential Information".

[54] Anthropic, "Building Safeguards for Claude". This post details Anthropic's classifier-based safety monitoring: fine-tuned Claude models ("classifiers") monitor conversations in real time, and a "hierarchical summarization" technique condenses individual interactions into account-level summaries for detecting aggregate-level concerns such as automated influence operations. Conversations triggering these classifiers may be escalated for human review and retained under the extended safety-retention schedule.

[55] See footnote 1.

ChatGPT Taskforce Report makes clear, an incomplete response that risks transferring compliance responsibility to users.

This is why this series proposes a **Sealed Mode** as a centrepiece privacy-by-design recommendation. The proposal is not that providers must guarantee absolute secrecy. It is that providers should offer at least one clearly labelled pathway in the consumer interface where the default architecture materially constrains downstream reuse and insider access, and where those constraints are legible and enforceable. Sealed Mode aims to shift from **promise-based privacy** (policy language that users must take on trust) to **constraint-based privacy** (technical and organisational limits that reduce what insiders and systems can do by default).

4.1 A practical design: an explicit sealed lane for high-stakes topics

A workable way to make Sealed Mode meaningful, without inviting universal "privacy laundering", is to implement it as an explicit lane for a limited set of high-stakes use cases. One concrete example would be a dedicated entry point such as "Health & Wellbeing — Sealed" (alternative labels could include "Care Mode", "Health Confidential", or "Protected Health Chat"). The interface would signal, in plain language, the enhanced protections and their limits: what is retained, for how long, and who can access it.

The health use case is a natural starting point for obvious reasons as this is an area where users constantly request advice to chatbots about extremely sensitive and confidential data. Health-related conversations are also a context where continuity has genuine value (a chatbot that remembers a user's conditions, medications, and allergies can give safer and more relevant answers), which means the design challenge is not to prohibit data use altogether but to confine it within auditable boundaries.

A sealed lane of this kind could combine the following default protections:

- **No training or model-improvement use** of sealed conversations (including fine-tuning and post-training alignment), subject only to narrow, defined safety exceptions.
- **Siloed, purpose-bound personalisation.** Sealed Mode does not require zero personalisation. In health contexts, continuity can be beneficial for safety and relevance. The design should therefore allow, with the user's explicit authorisation, a siloed "Health Profile" (for example, age range, conditions, medications, allergies) that is strictly confined to the sealed lane, excluded from training and advertising, and never used to personalise the general chatbot experience. Such a profile must be readily viewable and deletable by the user at any time.
- **No advertising surfaces** in the sealed lane, and no use of sealed-lane content as an advertising or ad-personalisation signal, even if the broader service supports ads elsewhere.
- **Strict retention bounds** and rapid deletion by default, with only a short, tightly scoped window for security and abuse-prevention logging.
- **Minimised routine human access:** sealed conversations are not eligible for random quality sampling; escalation to human review is exceptional, criteria-driven, and auditable.

- **Stronger access governance:** least-privilege controls, just-in-time access, immutable audit logs, and periodic independent review of access patterns.

Crucially, a sealed lane also creates an additional governance benefit: it makes it harder for providers to blur purposes under generic "improvement" language. The product is forced to articulate a clearer contract. Users receive tangible protection in exchange for accepting that the sealed lane is narrower, more tightly governed, and not designed for every use case.

4.2 Sealed Mode versus Temporary Chats: a different promise, a different architecture

Sealed Mode should not be conflated with existing "temporary" or "incognito" chat features. Temporary chats are a valuable minimisation tool, but they typically aim to reduce user-facing persistence (keeping a conversation out of chat history) and to limit some downstream reuse. They are generic and designed for any conversation a user prefers not to save.

Sealed Mode is different in at least three ways.

First, purpose and risk calibration. Temporary chat is a general privacy hygiene feature. Sealed Mode is designed for a defined set of high-stakes disclosures where the harm from exposure is unusually severe and where users reasonably expect higher confidentiality.

Second, insider-access constraints. Temporary chat modes rarely promise that content is structurally insulated from routine sampling or operational access. Sealed Mode targets precisely that internal boundary: no routine sampling, sharply limited human access, and strict escalation rules.

Third, continuity without leakage. Temporary chats are intentionally disposable, which makes them ill-suited for health contexts where continuity is beneficial. Sealed Mode can support continuity through siloed, user-controlled health profiles and narrowly bounded retention, without allowing sensitive information to leak into general memory, training, or monetisation.

In short, Temporary Chat answers the question "can I avoid saving this conversation in my sidebar". Sealed Mode answers a different and deeper question: "can I safely use this system for high-stakes disclosures without my conversation travelling through multiple systems and human hands by default".

4.3 Feasibility and abuse: designing constraints without disabling safety

A predictable objection is that if users can opt into Sealed Mode, they will use it for everything, including abusive or illegal content, thereby undermining safety monitoring. The objection is serious but does not make the concept infeasible. It means Sealed Mode must be designed as a *constrained* lane with complementary safeguards.

First, **topic-scoped functionality.** In a "Health & Wellbeing — Sealed Mode" lane, the system can be configured to answer only within a health-and-wellbeing domain and to refuse or redirect unrelated requests. The point is not perfect classification, but a strong default that makes "sealed" an honest label rather than a general-purpose cloak.

Second, **safety controls under enhanced safeguards.** Sealed Mode can preserve safety by allowing narrowly bounded automated checks (for example, for imminent self-harm

threats or certain abuse categories) and by enabling exceptional escalation pathways when a high-severity trigger is met. The key is that escalation should be rare, criteria-driven, documented, and auditable, not routine sampling. The goal is not to guarantee anonymity from the law or to enable unlawful behaviour; it is to reduce routine exposure for users with the highest confidentiality interests.

4.4 Cryptographic and infrastructural protections: from promises to constraints

A core challenge for Sealed Mode is credibility. The more sensitive the conversation, the less users should be asked to rely on a mere policy promise. This is where cryptographic and infrastructure-level protections become relevant. They will not eliminate all risk, and they raise genuine tradeoffs with safety monitoring and product functionality, but they can materially reduce both insider access and the blast radius of external exposure.

At minimum, Sealed Mode should be designed around segmentation and hardening: sealed-lane transcripts stored in a logically separate environment, encrypted at rest with strong key management, with strict access controls and audit logs. Beyond that baseline, providers should explore a spectrum of architectures that reduce the practical ability of insiders to access plaintext content: client-side encryption for stored transcripts, hardware-backed secure enclaves or confidential computing for processing, and minimisation of plaintext retention through short-lived processing and rapid deletion.

The feasibility of such approaches is no longer speculative. Apple's Private Cloud Compute, launched in 2024, demonstrates that consumer-scale AI inference can be processed within hardware-backed secure enclaves, where data is never stored after processing and the entire software stack is subject to independent verification. Meta's Private Processing for WhatsApp, launched in 2025 using AMD SEV confidential computing, allows users to invoke AI features on their messages within an environment where no one, including WhatsApp itself, can access the underlying content.[56] These are not research prototypes; they are production systems operating at consumer scale. The EDPB's own April 2025 commissioned-report on AI privacy risks and mitigations in LLMs identifies encryption, federated learning, and differential privacy as applicable mitigations for the data-processing risks inherent in large language model deployments.[57] And the confidential computing market is maturing rapidly, with Gartner projecting that by 2029 over 75% of processing operations in untrusted infrastructure will be secured by confidential computing.[58]

These measures also have implications for Part II of this series. Strong encryption, segmentation, and retention minimisation can reduce breach impact and narrow the scope of data available for disclosure under civil litigation. Sealed Mode could therefore

---

[56] In 2025, Meta launched Private Processing for WhatsApp, allowing users to invoke AI features on their messages within a confidential computing environment powered by AMD SEV, where no one, including WhatsApp, can access the underlying messages. See AMD, Your AI Provider Doesn't Need to See Your Data to Work on It (December 2025). Apple's Private Cloud Compute (PCC), launched in 2024, provides a similar architecture for Apple Intelligence: inference is processed in hardware-backed secure enclaves, data is never stored after processing (stateless computation), and the entire software stack is subject to independent verification. See Apple Security Research, Private Cloud Compute: A New Frontier for AI Privacy in the Cloud (2024). These industry examples demonstrate that constraint-based privacy for cloud AI processing is not speculative; it is already being deployed at consumer scale.

[57] Isabelle Barbera (commissioned by EDPB), "AI Privacy Risks & Mitigations – Large Language Models (LLMs)" (April 2025), addressing encryption, federated learning, and differential privacy as applicable mitigations for LLM data-processing risks. See also On the concept of "privacy by design" as applicable to consumer chatbot interfaces, see EDPB Guidelines 4/2019 on Article 25 Data Protection by Design and by Default, Version 2.0 (adopted 20 October 2020), paras. 69–70.

[58] Gartner's Top Strategic Technology Trends for 2026 (October 2025).

be designed not only as a "no training, less review" lane, but also as a resilience measure against the external exposure vectors that Part II will examine.

4.5 How Sealed Mode could be overseen

A legitimate question is how Sealed Mode would be overseen. The options range across a wide spectrum: from self-declaratory commitments by providers (analogous to how companies currently represent that they offer end-to-end encryption), through voluntary technical standards or certification frameworks (such as an ISO-type standard for high-assurance conversational privacy), to formal regulatory seals or certification mechanisms overseen by data protection authorities or sector-specific regulators. Each approach carries trade-offs. Self-declaration is the fastest to deploy but the hardest to verify independently. Technical standards offer interoperability and auditability but take time to develop and risk rigidity in a fast-moving field. Regulatory certification provides the strongest credibility signal but may be impractical if it cannot keep pace with product evolution. This study does not prescribe a single governance model. The more important point is that some form of independent verifiability is essential if Sealed Mode is to be more than a marketing label. Providers are best placed to propose architectures that are both effective and operationally realistic; regulators are best placed to define the minimum assurances that any such mode must deliver and to determine whether existing certification mechanisms (such as those contemplated under Articles 42 and 43 GDPR, or the conformity assessment procedures under the AI Act) can be adapted to serve this purpose. What matters is that the answer is found, not that it is found in a single predetermined form.

4.6 Scope and responsibility: this study proposes the principle and design objective, not a one-size-fits-all blueprint

This paper does not claim to provide the single perfect implementation of Sealed Mode. That is neither realistic nor necessary. The central claim is more modest and more urgent: the current consumer privacy toolkit is not commensurate with the way chatbots are actually used at scale for medical, mental-health, and crisis-adjacent disclosures. The gap between user expectations and actual protections is too wide, and it is growing as these systems become more fluent, more continuous, and more embedded in daily life.

Stakeholders should treat Sealed Mode, or an equivalent high-assurance pathway, as a priority design and governance problem. Providers are best placed to determine the most workable architecture[59]: the right balance of encryption, segmentation, topic scoping, and safety overrides will depend on each provider's infrastructure, threat model, and regulatory environment. But the obligation to pursue it is shared across product teams, security engineers, privacy professionals, regulators, and the research community. Health was chosen here as the illustrative use case because the regulatory gap and the disclosure harm are both well documented, but the same logic extends in principle to legal consultations, crisis-adjacent conversations, and other contexts where the consequences of disclosure are severe and the user's expectation of confidentiality is most acute.

[59] Several recommendations in this paper (including decoupling chat history from training, strengthening regurgitation safeguards, and implementing tiered confidentiality pathways) may require non-trivial engineering choices and may entail trade-offs in cost, latency, feature availability, or, in some settings, model quality. This paper advances the governance principle and design objective: where such trade-offs exist, they should be minimized and made explicit, and users should not have to 'pay' for privacy through hidden usability penalties.

5. Conclusions and recommendations

Human review is not a marginal exception; it is a structural feature of consumer chatbot governance. The key differences across providers are (i) how candidly review is disclosed at the interface level, (ii) what purposes are invoked, (iii) whether retention practices amplify the exposure, and (iv) whether real-world incidents have demonstrated the gap between policy assurances and actual outcomes. Gemini sets a benchmark for interface honesty, but its disclosures also show that the default experience can involve human review and three-year retention of reviewed chats, with no user-side mechanism to delete reviewed copies. Other providers rely more heavily on help pages and policy text, which weakens real-world transparency for typical users. Finally, it must be emphasized that human review is not the only way confidentiality can fail; interface-level design choices (sharing, export, collaboration) can expose chats at scale, a risk examined in Part II of this series.

Recommendations

Recommendation 4 (standardized interface disclosure). Providers should adopt a clear, standardized notice in the interface stating that a subset of chats may be reviewed by humans for safety, quality, and support, and clarifying the retention posture for reviewed chats. The Gemini warning is a useful model, but under GDPR fairness principles it must be understood as a transparency measure, not as a transfer of responsibility to the user.

Recommendation 5 (event-based transparency). Where feasible, providers should notify users when a conversation is escalated for human review, or at least provide a user-visible audit log indicating access categories and retention consequences.

Recommendation 6 (Sealed Mode for high-stakes topics). Providers should consider offering at least one clearly labelled, consumer-facing sealed lane for high-stakes use cases (for example, "Health & Wellbeing — Sealed Mode"), with strict defaults: no training, siloed purpose-bound personalisation, no ads, strict retention limits, and sharply limited human access governed by narrow, auditable safety exceptions. Providers should complement these policy rules, to the extent possible, with technical and infrastructural measures, including encryption, segmentation, and where feasible confidential computing, that move from promise-based privacy to constraint-based privacy. The aim is not to prescribe a single architecture but to establish the principle that the most sensitive consumer conversations deserve protections commensurate with the trust that users place in them.

III. Monetizing the Confessional: Ads, Profiling, Commercialization

1. Introduction: what is being measured, and why it is the hardest trust test

This section addresses the most fragile dimension of consumer chatbot trust: commercialization. A user may accept that a provider stores chats for functionality, or even that it reviews a subset for safety. But the moment conversations are linked to revenue extraction, the relationship changes. Advertising, profiling, and other monetization pathways create an appearance (and sometimes a reality) of conflicting incentives: the provider's business interest in maximizing engagement and conversion can sit uneasily alongside the user's expectation of discretion.

The shift is not hypothetical. In May 2024, OpenAI CEO Sam Altman described the prospect of advertising in AI chatbots as "uniquely unsettling" and called it "a last resort for us for a business model".[60] Twenty months later, in January 2026, OpenAI announced it would begin testing ads in ChatGPT for Free and Go tier users in the United States, and the test launched on 9 February 2026.[61] The speed of this reversal illustrates how rapidly the economics of large-language-model inference can overwrite even a CEO's stated preferences. OpenAI lost close to $8 billion in 2025, and only about 5% of its roughly 800 million weekly active users pay for subscriptions.[62]

The research questions this section pursues are concrete and time-sensitive. Are ads currently shown in the interface, and if so where? Are ads selected purely from the current thread (contextual placement), or can past chats, memory, or other signals be used (behavioral personalization)? Can users opt out of ads entirely or of personalization, and how? Finally, even where ads are absent today, do providers disclose other monetization uses of chat content, or does the broader business model create residual risks that users should reasonably consider?

2. Comparative table: current state of affairs (February 2026)

Table 3 compares the five providers across four dimensions: whether ads are currently shown in the interface (and where they are excluded); whether ad selection draws on user interactions beyond the current thread; whether users can opt out of ads or of personalization; and what residual monetization risks exist even where ads are not currently deployed. Primary sources are the providers' own documentation and public statements.

---

[60] Sam Altman, remarks at Harvard University event, May 2024: "Ads plus AI is sort of uniquely unsettling to me. I kind of think of ads as a last resort for us for a business model". Altman added that he would only pursue advertising "if that was the only way to get everyone in the world access to great services". See Men's Journal, ChatGPT Moves Forward With CEO's 'Last Resort' (17 January 2026). On an earlier occasion, Altman stated: "I kind of hate ads as an aesthetic choice". See Beebom, ChatGPT is Getting Ads Despite Sam Altman Calling It 'Uniquely Unsettling' (17 January 2026).

[61] *OpenAI, Our approach to advertising and expanding access* (16 January 2026) (contextual ads at bottom of answers; principles; ChatGPT Go launch). See also OpenAI, Testing ads in ChatGPT (9 February 2026) (start of U.S. test for Free/Go tiers; ad-free premium tiers; privacy commitments).

[62] Financial pressure is indeed big part of the context. OpenAI lost close to $8 billion in 2025, with the Financial Times describing it as an "era-defining money furnace".'Of ChatGPT's roughly 800 million weekly active users, only about 5% pay for subscriptions. OpenAI's infrastructure commitments total approximately $1.4 trillion over the next eight years. See PC Gamer, Here we go: OpenAI CEO Sam Altman once called it a 'last resort' (16 January 2026); Tony Grayson, OpenAI Advertising in ChatGPT (January 2026) (citing J.P. Morgan analysis that the AI industry needs $650 billion in annual revenue to deliver a 10% return on infrastructure investments through 2030).

Table 3: Commercialization of chat data: ads, targeting, opt-outs, and residual risks (as of 23/02/26)

Company	Ads shown?	Informed by user interactions?	Opt-out	Residual risks
OpenAI (ChatGPT)	YES. Limited U.S. test for logged-in adults on Free and Go tiers. Ads appear at bottom of answers, labeled and separated. No ads on Plus/Pro/Business/Enterprise/Education. No ads in Temporary Chats or ChatGPT Atlas browser.	For US pilot: Personalization enabled by default. When on: past chats, memory, and ad interaction signals may inform ad selection. When off: contextual matching to current thread only.	Settings > Ad Controls: toggle personalization off; dismiss ads; clear ad data. To avoid ads entirely: upgrade to ad-free plan or choose free option with additional limits.	OpenAI commits ads do not influence model answers and advertisers do not receive chats or memories. Residual risk: internal profiling if personalization remains on by default; geographic expansion triggers EU law scrutiny.
Google (Gemini)	NO ads in Gemini Apps (as disclosed). Google DeepMind CEO confirmed 'no plans whatsoever' (Davos, January 2026).	Google states Gemini conversations are not used to show ads.	N/A for ads. Users can control Activity and choose Temporary Chat.	Residual risk is ecosystem-based: Google's ad business makes the stated separation between Gemini chats and ad targeting consequential for trust.
Anthropic (Claude)	NO ads. Anthropic publicly positions Claude as ad-free ("Claude will remain ad-free").	N/A (no ads).	N/A (no ads).	Business model is subscriptions and enterprise contracts. Anthropic includes a narrow caveat: "Should we need to revisit this approach, we will be transparent about our reasons for doing so".
xAI / X (Grok)	No ads disclosed inside Grok chat itself. X as a platform is ad-supported and may show ads elsewhere. Elon Musk announced in August 2025 that he "plans to introduce ads to X's AI chatbot".	No ad selection logic for Grok chats is disclosed. X/xAI describe training uses of interactions.	N/A for in-chat ads. Users can opt out of training uses via X privacy setting; Private Chat not used for training.	Residual risk is high by business-model proximity: Grok is embedded in an ad-driven platform, creating structural incentives.
DeepSeek (DeepSeek Chat)	No ads disclosed in primary materials reviewed.	No ad targeting practices disclosed.	N/A for ads.	Residual risk is opacity: limited disclosures about commercialization pathways. Absence of disclosure is not the same as a commitment.

3. Analysis: why ads are a qualitatively different privacy boundary

Commercialization is a trust threshold. When monetization enters the chat window, users inevitably ask whether the system is still aligned with their interests. This is why, even before ads were deployed, OpenAI treated the topic as exceptional in its public communications, emphasizing that ads would be separated from answers and would not influence what the model says.

3.1 OpenAI's 2026 ad test[63]

The trajectory from stated reluctance to deployment is worth tracing, because it illustrates how quickly structural pressures can reshape privacy preferences. In May 2024, Altman described AI-plus-ads as "uniquely unsettling". In June 2024, he softened: "I'm not totally against it. I can point to areas where I like ads. I think ads on Instagram, kinda cool".[64] By February 2026, the Instagram model had become the explicit aspiration for ChatGPT ads: discovery-oriented placements that help users "find stuff they like that they otherwise never would have".[65] The gap between those statements and the deployment is narrower than the rhetoric suggests.

OpenAI's current approach is documented in three layers. First, on 16 January 2026, OpenAI set out its advertising principles and announced an intention to test ads in the United States for logged-in adults on Free and Go tiers, starting with ads shown at the bottom of answers when there is a relevant sponsored product or service based on the current conversation.[66]

Second, on 9 February 2026, OpenAI announced the start of that test and reiterated two commitments that go directly to privacy and integrity: (i) ads do not influence ChatGPT's answers, and (ii) conversations are kept private from advertisers, who receive only aggregate performance information.[67]

Third, OpenAI's Help Center[68] provides operational detail. Ads are limited to Free and Go users in the U.S.; ad controls are available in Settings; Temporary Chats do not show ads; and sensitive or regulated topics are excluded. The Help Center also clarifies an important distinction between contextual relevance and behavioral personalization. However, a

[63] While Perplexity is not included in the comparative analysis in this paper (for the reasons explained in the Introduction), it is worth noting that Perplexity was an early and prominent "answer engine" to experiment with advertising in a generative-AI search interface. In late 2024, Perplexity announced that ads would be introduced in the U.S. and formatted as "sponsored follow-up questions" (and related paid placements), clearly labelled as "sponsored", while emphasizing that answers to sponsored questions remain generated by Perplexity's AI and are not written or edited by the sponsoring brands. For context also, GenAI advertising formats predate Perplexity in other consumer chatbot environments, such as Snapchat's My AI, which incorporated "Sponsored Links" in partnership with Microsoft in late 2023.

[64] In June 2024, on OpenAI's podcast, Altman softened further: 'I'm not totally against it. I can point to areas where I like ads. I think ads on Instagram, kinda cool.' By January 2026, he framed the Instagram model as the aspiration for ChatGPT ads: 'An example of ads I like are on Instagram, where I've found stuff I like that I otherwise never would have.' The shift from 'uniquely unsettling' to 'kinda cool' took approximately twenty months. See AOL, ChatGPT is getting ads. Sam Altman once called them a 'last resort' (16 January 2026).

[65] See Sam Altman's interview with CNBC, February 19, 2026.

[66] OpenAI, Our approach to advertising and expanding access (16 January 2026).

[67] OpenAI, Testing ads in ChatGPT (9 February 2026) (start of U.S. test for Free/Go tiers; ad-free premium tiers; privacy commitments).

[68] OpenAI Help Center, Ads in ChatGPT (placement; controls; personalization signals; exclusions).

critical detail deserves emphasis: ad personalization is enabled by default.[69] In the U.S. Pilot, when personalization is on, OpenAI states it may use additional signals including past chats, memory, and ad interaction history to select ads. Users who wish to limit ad selection to the current thread must affirmatively navigate to Settings and toggle personalization off. The default matters because it determines the baseline experience for the large majority of users who never adjust settings.

Early data from the rollout reveals further details. The initial beta carries a $60 CPM rate and a $200,000 minimum advertiser commitment, limiting early access to large-budget brands such as Expedia, Target, Ford, and Adobe.[70]

OpenAI's CEO of Applications, Fidji Simo, formerly CEO of Instacart where she built the company's advertising platform, has stated that OpenAI would be "extremely respectful" of user data and that "the thing people don't like about ads very often is not the ads themselves, it's the use of the data behind the ads".[71] That observation points to exactly the right question. But the answer depends on whether the default architecture matches the stated principle.

3.2 Competitor positioning: Anthropic's ad-free identity and the Super Bowl moment

Anthropic has responded by making "no ads" a defining feature of Claude's identity. In a blog post published on 4 February 2026, it argued that even ads displayed separately inside the chat window would compromise what Claude is meant to be: a clear space to think and work.[72] This is a strategic positioning move, but it also functions as a normative claim about what is presented as the tricky relations between conversational intimacy and ad monetization.

The positioning was amplified through an $8 million Super Bowl campaign.[73] Spots titled "Violation"[74], "Betrayal"[75], and "Deception"[76] dramatized the scenario of an AI assistant pivoting mid-conversation from sensitive personal advice to a jarring product plug, closing with the tagline: "Ads are coming to AI, but not Claude". Anthropic also disclosed that Claude Code and Cowork had already generated at least $1 billion in annualized revenue through enterprise partnerships, underscoring that an ad-free model can be financially viable at scale. The company did, however, include a narrow caveat in its blog post: "Should we need to revisit this approach, we will be transparent about our reasons for doing so". The conditional framing is notable. It preserves optionality while placing the current commitment on the record.

---

[69] By default, ad personalization is enabled for eligible users. When turned on, OpenAI states that ads may be personalized "based on your chats and any context ChatGPT uses to respond to you including ad interaction history and, where the 'Past chats and memory' toggle remains active, stored memories and prior conversations. See OpenAI Help Center, Ads in ChatGPT. The distinction between contextual and behavioral ad selection is legally consequential: contextual matching against the current thread involves a different category of processing from longitudinal profiling across past chats and stored memories.

[70] See Winbuzzer, ChatGPT Ads Now Hit on Your Very First Message (21 February 2026). Notably, sponsored placements can appear after a user's very first message, before any conversational context has been established.

[71] See AOL, ChatGPT is getting ads (16 January 2026).

[72] *Anthropic, Claude is a space to think* (4 February 2026). The company's analysis of user conversations found that "many involve sensitive personal topics, complex engineering tasks, or deep work where ads would feel inappropriate".

[73] See Technobezz, Anthropic spends $8 million on Super Bowl ads (5 February 2026).

[74] Antropic, "Can I get a six pack quickly?", You Tube.

[75] Antropic, "How can I communicate better with my mom?", You Tube.

[76] Antropic, "What do you think of my business idea?", You Tube.

Google's positioning adds a further data point. DeepMind CEO Demis Hassabis confirmed at the World Economic Forum in Davos in January 2026 that Google has "no plans whatsoever" to introduce ads into Gemini. His reasoning was explicitly trust-based: "If you want a truly universal assistant that you can trust, that is personally tailored to you and knows a lot about you, then you want to be sure that the recommendations it gives you are really good for you, unbiased and unadulterated".[77] Hassabis also commented that he was "a little bit surprised" at how quickly OpenAI moved on ads, suggesting it might reflect "revenue pressure". Google can afford this stance because its advertising revenues from search and other services cross-subsidize AI development. The question is whether that structural advantage will hold as inference costs grow.[78]

Grok (xAI/X) is also moving toward an advertising-adjacent model. In a live discussion with advertisers reported in early August 2025, Elon Musk indicated that X would allow marketers to pay to appear in Grok's suggestions, framing this as a way to monetize the very moment a user is trying to "solve a problem" and to fund the compute costs required to run large models. Reporting also described broader ambitions to automate parts of X's ad pipeline and improve targeting and relevance using xAI's technology, against a backdrop of X's continuing effort to revive its advertising business and ongoing concerns from some advertisers about brand safety on the platform.[79]

3.3 The Altman-Anthropic clash and the governance questions it surfaces

The immediate trigger for the advertising debate in early 2026 was not an abstract policy document, but a public clash. Anthropic used Super Bowl advertising to mock the idea of ads in chatbot conversations. Sam Altman responded on X with unusual directness: he called the campaign "clearly dishonest", described Anthropic as serving "an expensive product to rich people", and emphasized the economic logic of an ad-supported tier as a way to keep extending free access to very large numbers of users who can't pay for subscriptions.[80]

This exchange matters because it exposes a core governance question. If ads are justified as the price of broad access, then privacy protection and interface design become the foundation of legitimacy. Users will tolerate monetization only if they can predict, with reasonable precision, what is and is not being used to select ads. That is why OpenAI's published ad rules and controls are not merely marketing language: they are boundary conditions for consumer trust.

As mentioned earlier, Altman has invoked an "Instagram-style" notion of discovery ads: ads that help users find something useful rather than ads that feel like surveillance. But the decisive distinction is not rhetorical. It is technical and legal: contextual placement based on the current thread is a different category of processing from behavioral personalization using past chats, memory, or inferred interests. OpenAI's documentation draws that line by making personalization a toggleable setting and by describing the

[77] See Axios, "Exclusive: DeepMind CEO 'surprised' OpenAI moved so fast on ads" (21 January 2026). Google Ads VP Dan Taylor had previously stated in December 2025 that "there are no ads in the Gemini app and there are no current plans to change that". See 9to5Google, "Google reiterates Gemini won't have ads" (21 January 2026),

[78] Advertising insiders report Google has discussed possible Gemini ad placements with media buyers. See Adweek (exclusive): "Google Tells Advertisers It'll Bring Ads to Gemini in 2026", December 8, 2025.

[79] See Financial Times, "Elon Musk to Intreoduce Ads to X's AI Chatbot", August 7, 2025.

[80] See Sam Altman's reaction on X. See also TechCrunch, Sam Altman got exceptionally testy over Claude Super Bowl ads (4 February 2026). Altman argued that OpenAI needed ads to "bring AI to billions of people who can't pay for subscriptions". See also Adweek, AI rivalry heats up ahead of Super Bowl (5 February 2026).

additional signals that may be used when personalization is enabled. But since personalization is enabled by default, the line is drawn in the wrong place for users who do not actively manage their settings.

3.4 A word of caution: the European regulatory landscape and the cost of unpreparedness

Any expansion of chat-based advertising into the EU would confront a layered and unusually protective regulatory framework. The details of that framework, and the specific compliance challenges it poses for conversational AI advertising, deserve a separate and dedicated analysis. But a word of caution is warranted here, grounded in recent experience.

The history of AI chatbot launches in Europe should serve as a cautionary tale. When OpenAI released ChatGPT in November 2022 with minimal GDPR safeguards, the result was an immediate regulatory clash: the Italian Garante imposed a temporary ban in March 2023 (the first in a Western country), the EDPB assembled a dedicated ChatGPT Taskforce, investigations multiplied across European jurisdictions, and OpenAI was forced to retrofit the privacy protections it should have built in from the start, eventually facing a EUR 15 million fine from the Garante in December 2024 for, among other things, processing personal data without an adequate legal basis and failing to comply with information obligations.[81] The same pattern repeated itself with DeepSeek's European launch in early 2025: the company appeared to enter the European market without incorporating any of the lessons from the ChatGPT episode, triggering a tsunami of reactions from data protection authorities across Europe and beyond, a ban by the Italian Garante within 72 hours, investigations in over thirteen European jurisdictions, and the creation by the EDPB of a dedicated AI Enforcement Task Force.[82]

The lesson from both episodes is consistent: launching a consumer AI service in Europe without having thoroughly studied and addressed the regulatory requirements leads to costly regulatory confrontation, reputational damage, and forced retroactive corrections. There is no reason to believe that deploying advertising within chatbot conversations would follow a different pattern. If anything, the regulatory risks are more acute, because ads in chatbots sit at the intersection of multiple overlapping EU instruments. The GDPR imposes lawful-basis, transparency, and purpose-limitation requirements that are particularly demanding when the data involved is intimate and routinely includes special categories (health, mental health, sexuality, religion, political opinions). The Digital Services Act's Article 26(3) flatly prohibits presenting advertisements based on profiling using special-category data, a prohibition that is absolute and cannot be overridden by consent. The ePrivacy Directive adds further constraints on the storage and access of information on terminal equipment. And the AI Act also includes several provisions that could be relevant.

Moreover, the specific question of personalized advertising in digital platforms has been the subject of years of regulatory and judicial battles in Europe, with Meta at the center.

---

[81] On OpenAI's ChatGPT launch and the ensuing regulatory confrontation, see the Italian Garante's decisions of 30 March 2023 (temporary restriction) and 20 December 2024 (EUR 15 million fine for, *inter alia*, absence of lawful basis for processing and failure to comply with information obligations). The EDPB ChatGPT Taskforce was established in April 2023 and published its report on 23 May 2024.

[82] On DeepSeek's European launch and the global regulatory response, see T. Christakis, DeepSeek and the China Data Question: Direct Collection, Open Source, and the Limits of Extraterritorial Enforcement, IAPP (February 2025), ; T. Christakis & P. Raj, DeepSeek One Year Later: Regulatory Storm, Global Surge, AI-Regulation.com (28 January 2026).

The CJEU's landmark judgment in Case C-252/21 (*Meta Platforms v. Bundeskartellamt*, July 2023) examined the conditions under which a dominant platform can lawfully process personal data for behavioral advertising, imposing stringent requirements on consent and necessity.[83] The EDPB subsequently issued an urgent binding decision in November 2023 ordering Meta to stop processing personal data for behavioral advertising on the basis of contractual necessity or legitimate interest, and its April 2024 Opinion on "consent or pay" models further constrained the options available to platforms seeking to fund services through personalized ads.[84] The European Commission then fined Meta under the Digital Markets Act in April 2025, finding that its "pay or consent" subscription model did not comply with Article 5(2) DMA because it failed to offer users an equivalent alternative using less personal data.[85] All this represents a concrete, evolving, and increasingly restrictive body of law and practice that any provider planning to introduce advertising into a conversational AI service in Europe will need to navigate with extreme care.

The recommendation, then, is straightforward: providers should study these questions with the seriousness they deserve *before* deployment, not after. The regulatory cost of unpreparedness, as ChatGPT's 2022 European launch and DeepSeek's 2025 launch both demonstrate, is far higher than the cost of getting it right from the beginning.

3.5 Residual risks in ad-adjacent ecosystems and the retention incentive

Grok illustrates a distinct commercialization risk. Even if Grok does not currently display ads in the chat interface, it operates inside an ad-supported social platform. In such environments, users may reasonably worry that conversational interactions could become a future signal for profiling, segmentation, or ad measurement, particularly when the same account identity is used across platform functions. The risk is not only what is stated today, but what the architecture enables tomorrow.[86] Where providers do not publish clear commitments excluding chat data from ad targeting, consumer trust depends on silence rather than enforceable representations.

A related structural concern applies across providers. Advertising incentives can increase pressure to retain conversational data for longer periods or to find new ways to extract value from it.[87] This creates a tension with the data-minimization and storage-limitation

---

[83] CJEU, Case C-252/21, *Meta Platforms Inc. and Others v. Bundeskartellamt*, ECLI:EU:C:2023:537 (4 July 2023). The Court held, *inter alia*, that while dominance does not as such preclude valid consent, it is 'an important factor in determining whether the consent was in fact validly and, in particular, freely given' (para. 149), and acknowledged the possibility of an 'appropriate fee' as an alternative to consent (para. 150).

[84] EDPB, *Urgent Binding Decision on Processing of Personal Data for Behavioural Advertising by Meta* (1 November 2023) (ordering Meta to cease processing for behavioral advertising on the basis of contractual necessity or legitimate interest within the EU/EEA); EDPB, Opinion 08/2024 on *Valid Consent in the Context of Consent or Pay Models Implemented by Large Online Platforms* (17 April 2024), finding that in most cases large online platforms cannot comply with GDPR consent requirements by offering only a binary choice between consenting to behavioral advertising or paying a fee.

[85] European Commission, Decision of 23 April 2025, finding Meta in breach of Article 5(2) DMA for failing to provide users with a 'less personalised' equivalent alternative that does not require payment. The Commission imposed a fine and required Meta to offer a free-of-charge option with reduced personalization.

[86] For clarity, this paper does not assume that a company's general advertising capabilities are necessarily integrated into its consumer chatbot product. For example, Google publicly states that Gemini conversations are not used to show ads, and, as a matter of product architecture, advertising functionality may be absent from the Gemini interface even within a company that operates a large advertising business. The point made here is narrower: as monetisation strategies evolve (and as some providers begin testing ads inside chat), user trust depends on clear, durable and verifiable boundaries around whether and how conversation data can influence commercial outcomes, rather than on user assumptions about a provider's business model.

[87] On advertising incentives increasing pressure to retain conversational data for longer periods, see Digital Watch Observatory, ChatGPT and the rising pressure to commercialise AI in 2026 (21 January 2026): 'Advertising incentives can increase pressure to store conversations for longer periods or to find new ways to extract value from them.'

principles. If conversations become an input for ad-selection algorithms, the economic incentive to preserve those conversations grows, and the case for aggressive deletion weakens. As shown earlier, empirical research has documented the high frequency with which users disclose health data, financial details, and professional secrets in chatbot conversations. When such conversations serve simultaneously as inputs for ad targeting, the privacy stakes of commercialization are compounded.[88]

4. Conclusions and recommendation

Commercialization is the sharpest test of chatbot privacy because it collapses the boundary between conversation and commerce. The current landscape is heterogeneous. Gemini states that chats are not used to show ads, and Google's leadership has repeatedly confirmed that position. Anthropic positions Claude as explicitly ad-free and has invested heavily in making that distinction visible to consumers. OpenAI is testing ads in a limited U.S. context with controls and exclusions, but with personalization enabled by default. Grok's embedding in an ad-supported platform creates structural residual risk even if ads are not currently displayed in-chat. DeepSeek's disclosures do not give consumers a reliable basis to assess ad-related risks.

This area is evolving quickly. If providers deploy ads more broadly, and especially in Europe, the sustainability of consumer trust will depend on whether commercialization is designed and governed in a way compatible with existing legal frameworks, starting with the layered architecture of EU law requirements. The scale of what is at stake should not be underestimated: OpenAI has reported more than 120 million average monthly active recipients in the EU for ChatGPT Search alone for the six-month period ending 30 September 2025.[89]

A large number of concrete recommendations could be formulated on how to design, govern, and disclose chat-based advertising. The subject warrants a separate, dedicated study. For present purposes, the key recommendation is methodological and institutional:

Recommendation 7 (Regulatory-first approach to conversational advertising): Before deploying or expanding conversational advertising, and particularly before any deployment in Europe, providers should map the full set of difficult questions that this practice raises, covering at minimum: the signals used for ad selection and their default settings; the handling of sensitive-category data that users routinely disclose in chats; retention periods for ad-related processing; the transparency mechanisms available in a conversational interface; the interaction between several relevant EU law instruments; and the auditability of the entire pipeline. Any provider that concludes, after this mapping exercise, that the regulatory and trust risks are manageable should adopt a package of protective measures sufficient to preserve user trust and to comply with an unusually protective regulatory framework. The burden of proof lies with the provider, not the user. What is at stake is not simply regulatory compliance, but the foundational relationship between a conversational AI and the people who confide in it.

---

[88] Open AI notes that "ads are not eligible to appear near sensitive or regulated topics, including health, mental health, or politics". There is no clear and explicit commitment, though, that sensitive data (and especially, for Europeans, Article 9 of the GDPR special categories of data) will not be used for profiling in general.

[89] Euronews, ChatGPT could face strictest set of EU rules as it hits 120 million users in Europe (22 October 2025) (OpenAI disclosed approximately 120.4 million average monthly active recipients in the EU for ChatGPT Search for the six-month period ending 30 September 2025).

IV. The Business Plumbing of a Chat: 'No Sale' Commitments, Vendor Sharing, Affiliates, and Ecosystem Spillover

1. Introduction: what is being measured, and what is intentionally out of scope

Sections I, II and III addressed three obvious controller-side uses of chat interactions: training, human review, and advertising-adjacent monetization. Yet even when a provider promises not to "sell" personal data, and even when it keeps ads out of the chat window, users may reasonably ask how widely their conversations circulate within the provider's operational infrastructure. This section examines that boundary: how chat-related data may be disclosed to vendors and contractors, transferred within corporate groups, and reused for business operations that are neither training nor advertising, including analytics, fraud and abuse prevention, customer support tooling, and infrastructure monitoring.

A critical distinction must be drawn at the outset. Operational sharing with downstream service providers (cloud hosts, safety auditors, customer support tooling) is legally and functionally different from selling data for a third party's independent commercial use. A "sale", as the term is understood in US state privacy laws and in ordinary language, implies that the recipient acquires the right to use the data for its own purposes. Operational sharing, by contrast, involves processors who act on the controller's instructions, typically under strict contractual confidentiality. Under the GDPR, Article 28 imposes specific obligations on processors: they may process personal data only on documented instructions from the controller, must implement appropriate security measures, and are subject to contractual and statutory limits on secondary use. This means that not all data flows carry the same privacy risk, and this section does not suggest otherwise. A provider that engages a vendor to perform a safety audit is taking a pro-privacy action, not extracting commercial value from user conversations. The question this section pursues is accordingly not whether operational sharing is improper — it is usually necessary and often serves users' interests — but whether consumers can realistically assess who may access their conversations, under what constraints, and for how long.

Two clarifications are also essential. First, this series is not about agentic AI or tool ecosystems in which the main chatbot is combined with third-party apps, plugins, extensions, or connected services that execute actions on a user's behalf. That topic raises distinct questions about downstream controllers, delegation, and transitive retention, and it deserves its own analysis.[90] Second, even within a "standalone consumer chatbot" scope, ecosystems matter. Some consumer chatbots are embedded in broader account systems and may, by design, share data with other services to fulfill a request. Accordingly, this section also assesses whether "combination with other services" is an opt-in feature, an opt-out feature, or a default architecture that users cannot realistically avoid.

2. Comparative table: "sale", sharing, and ecosystem connections (consumer use)

The table below compares the five major providers across four dimensions: whether they state they do not sell personal data; whether they disclose sharing with vendors and

[90] See for instance recent guidelines by See *Agencia Española de Protección de Datos* ('AEPD'), "Agentic AI From the Perspective of Data Protection", February 2026, 71 p.

contractors for operations; whether chat data may be shared with affiliates or related companies; and whether the chatbot is standalone by default or combined by design with other services. The purpose is to map the scope of access that each provider's disclosures imply.

Table 4 : "Sale", sharing, and ecosystem connections (as of 23 February 2026)

Provider	'No sale' / no sharing for ads?	Sharing with vendors / contractors?	Sharing with affiliates / related companies?	Standalone or combined by design?
ChatGPT (OpenAI)	YES. States it does not 'sell' Personal Data or 'share' for cross-context behavioral advertising.	YES. Uses vendors and service providers; processing may occur where vendors/providers are located.	YES. Processing/storage may involve affiliates and partners; internal transfers implied for operations.	Mostly standalone (chat experience within OpenAI account). Tool-like features can create downstream sharing, but this article brackets those.
Gemini (Google)	No single 'no sale' statement in Gemini hub; But clear 'no sale' commitment in Google's general safety/privacy practices policy.	YES. Google's general model relies on vendors; Gemini disclosures emphasize reviewer workflows and operational processing.	YES. Gemini can share data with other Google services to fulfil requests; those services then use data under Google Privacy Policy.	More combined by design. Gemini may share data (e.g., location, emails, photos) with Google services like Maps, Gmail, and YouTube; some Connected Apps are designed to work automatically.
Claude (Anthropic)	YES: states it does not use personal data to build profiles to sell/market, and does not sell the information to third parties.	YES. Discloses service providers and operational uses; contractors may be involved under controls.	YES. Standard policy structure includes affiliates/related entities for operations (where applicable).	More standalone. Consumer Claude operates within Anthropic's service; integrations are separate and out of scope here.
Grok (xAI / X)	YES (xAI policy): states it does not sell personal information or share for targeted/cross-context ads; does not use it for marketing.	YES. Discloses contracted service providers for hosting, cloud, analytics, support, safety monitoring, payments, etc.	YES. Discloses sharing with related companies where necessary for operations. X may share X data and Grok interactions with xAI.	Combined by architecture for many users: Grok on X is governed by X policies; xAI acquired X in March 2025, further integrating both entities.
DeepSeek Chat	YES (DeepSeek policy): states it does not sell personal data, does not use data for targeted ads, and does not share conversations with other users.	YES. Service providers (hosting/analytics/payments) under confidentiality agreements.	Unclear. Some policy versions do not detail affiliate transfers; disclosures are less operationally granular than others.	Mostly standalone as disclosed; however, ecosystem linkages are not described in sufficient detail to assess defaults.

3. Analysis: "what "no sale" does and does not tell users about confidentiality

The first lesson is that "no sale" commitments and operational data governance are distinct questions, and both matter. All five providers reviewed make some version of a "no sale" commitment. These commitments are genuine and important: they mean that the provider does not transfer user data to third parties for those third parties' own independent commercial purposes. A user can take meaningful reassurance from them.

But a "no sale" statement, by itself, does not describe the full scope of who may access conversation data within the provider's operational supply chain, or under what constraints. A provider that maintains a strict no-sale posture will typically still share chat-related data with a range of processors and sub-processors: cloud infrastructure, content delivery, analytics services, support ticket systems, safety operations, and human-review vendors. In the GDPR framework, these recipients act as processors under Article 28, bound to process data only on the controller's documented instructions. The contractual and statutory limits that apply to processors are significant safeguards, and they make operational sharing categorically different from a data sale in terms of the recipient's legal agency over the data.

The practical question, however, is whether the existence of these legal safeguards is sufficient on its own to sustain user trust, or whether it also matters that users can understand and verify who accesses their conversations. From the user's vantage point, the relevant question is not the legal characterisation of the data flow (sale versus processing) but something more immediate: can someone I do not expect read this conversation, and do I have any way of knowing? The processor framework under GDPR provides a strong legal basis for limiting secondary use, but it does not by itself ensure that users understand the practical scope of access, and it is that practical understanding that determines whether a user will confide.

A further precision is important. The "no-sale" commitments encountered in provider documentation are overwhelmingly framed in terms of US state privacy law, particularly the CCPA/CPRA definitions of "sale" and "cross-context behavioral advertising". These are meaningful guardrails within their statutory scope. In Europe, the relevant legal framework is different: the GDPR's controller–processor structure (Articles 26–28), combined with purpose limitation (Article 5(1)(b)) and the obligation to implement appropriate safeguards by design and by default (Article 25), provides a more granular governance architecture for operational data flows. This study does not conflate the two frameworks. Its point is narrower: regardless of which legal regime applies, users benefit when providers disclose, in consumer-facing language, how widely their conversations circulate and under what safeguards — not only whether the data is sold.

3.1 The operational transparency gap: vendors, contractors, and tooling

Operational sharing with vendors and contractors is not inherently a privacy concern. Many of these engagements — safety audits, abuse-detection outsourcing, infrastructure monitoring — are pro-privacy actions designed to protect users and the ecosystem. The processors involved are typically bound by contractual confidentiality, restricted to acting on the controller's instructions, and subject to the GDPR's processor obligations where European law applies. The privacy question is therefore not whether this sharing occurs, but how transparent providers are about it.

On this transparency dimension, a gap persists. Policies often list categories of recipients ("service providers", "contractors", "vendors") without identifying which categories ever see plaintext chat content, what safeguards apply, and how long operational logs persist. Gemini's materials are relatively candid about human review and long retention for reviewed chats, but the broader question of which internal and vendor systems can access conversation content remains difficult for ordinary users to infer. DeepSeek's no-sale and no-targeted-ad commitments are helpful, but the operational detail is thinner, which makes it harder to assess the practical scope of access.

The gap matters because conversations with consumer chatbots routinely contain material that users treat as confidential. Even where every link in the processing chain is contractually bound and GDPR-compliant, the user's practical sense of confidentiality depends on knowing that such safeguards exist and understanding their scope. Transparency about the chain of access controls, contractual obligations, and retention schedules is what allows users to form reasonable expectations — and it is the absence of that transparency, rather than the existence of operational sharing itself, that creates a trust deficit.

3.2 Ecosystem spillover: when "standalone" is not really standalone

The second lesson is architectural. Gemini and Grok illustrate that consumer chatbot experiences can be embedded in broader ecosystems where "sharing" is a feature, not an exception.

Gemini may share data with other Google services to fulfill a request. Its Privacy Hub[91] states, for instance, that it may share precise location data with Google Maps. Through the Connected Apps framework, Gemini can also interact with Gmail, Google Docs, Drive, Photos, YouTube, Calendar, and third-party applications such as Spotify and WhatsApp, and some of these apps are designed to work automatically with Gemini. Google's own Connected Apps personalization page states candidly that data shared between Gemini and Connected Apps "may relate to topics you find sensitive, like race, religion, and health, or confidential info", and that such data "is used to create personal insights about you, your relationships, and the people and things in your world".[92] In these designs, the meaningful question is not only what Gemini does, but what the receiving service does under its own policy framework.

A nuance is worth noting: ecosystem sharing of the kind described above is structurally different from the processor-based operational sharing discussed in Section 3.1. Where a chatbot shares data with another product or service that processes it for its own purposes (rather than solely on the chatbot provider's instructions), the receiving service may be acting as a separate controller, or as a joint controller, rather than as a processor. In these situations, the legal protections that apply to processor relationships (strict contractual limitation to the controller's instructions, prohibition of secondary use) may not fully apply. Users' ability to understand and control these cross-service data flows is accordingly a distinct governance challenge.

Grok on X presents an even more direct case of architectural integration. X may share X data, including Grok interactions and outputs, with xAI, and the governing policy depends on whether the user accesses Grok through X or through xAI's own app. Since xAI's

[91] Google, "Use & manage Connected Apps in Gemini".
[92] Google, About personalization with Connected Apps.

acquisition of X in March 2025, the two entities are further integrated, making the boundary between social media platform data and AI chatbot data increasingly porous.

3.3 The litigation retention problem

Even where a provider's own retention policies are reasonable, external legal processes can override them. I will analyse this issue in Part II of this series.

3.4 Questions that require more transparency and clarification

Several concrete transparency gaps recur across providers:

- Which categories of vendors and internal teams can access plaintext chat content, and under what access controls (least privilege, just-in-time access, dual control)?
- Are chat contents used in product analytics and A/B testing, and if so are they transformed (aggregation, redaction, pseudonymization) before analysts see them?
- Are chat transcripts exported into customer support systems when users request support, and what retention schedules apply there?
- When a chatbot shares data with another service to fulfill a request, is the user clearly told at the moment of sharing, and can the user disable that inter-service sharing without losing core functionality?

4. Conclusions and recommendations

This section has examined two distinct dimensions of the confidentiality boundary: operational sharing with processors and vendors, and ecosystem integration with other services. On the first dimension, the study finds that operational sharing is not a privacy failure: it is an inevitable and often privacy-protective feature of modern service delivery, governed in Europe by the GDPR's processor framework. The real concern is transparency: providers can make strong, public no-sale commitments and comply with their processor obligations, yet still leave users unable to assess who may access their conversations and under what constraints. On the second dimension, the study finds that for some providers, the chatbot is not a standalone service but part of a broader ecosystem where data flows to other products by design, sometimes involving cross-service sharing that goes beyond a simple controller–processor relationship.

Lessons learned

"No sale" commitments and operational governance are both important but serve different functions. A no-sale commitment assures users that their data is not transferred for third parties' independent commercial use. Operational governance (contractual processor controls, access limitations, logging, and audits) determines how securely data is handled within the provider's own supply chain. Both contribute to user trust, but neither alone gives users a complete picture of the confidentiality boundary.

Operational sharing can and should be disclosed more precisely, even where it is legally compliant. Users deserve to know which categories of recipients can see plaintext content and under what conditions, not because such sharing is improper, but because opacity

about the scope of access — however well-governed in law — erodes the practical sense of confidentiality that users rely on when deciding what to disclose.

Ecosystems make privacy boundaries dynamic. What starts as a chat may become an inter-service transaction, as the Gemini Connected Apps framework and the Grok/X integration illustrate.

Recommendations

Recommendation 8 (recipient transparency). Providers should describe, in consumer-facing language, which categories of vendors and internal teams can access plaintext chat content, and which receive only derived or redacted data. Where operational sharing serves a privacy-protective function, such as safety auditing, abuse detection, or security monitoring, providers should explain this purpose in terms that allow users to understand that such processing is undertaken in their interest and is subject to contractual and statutory constraints on secondary use.

Recommendation 9 (ecosystem boundary notices). Where a chatbot shares user data with another service to fulfill a request, the interface should disclose that sharing at the moment it occurs and provide a practical control to disable it.

Recommendation 10 (separate "operations" from "improvement"). Providers should avoid collapsing broad operational uses into "improvement" language, and should separate product analytics, safety operations, and training in user controls and disclosures.

General Conclusions

Mapping the Internal Boundary of Consumer Chatbot Privacy

Every day, hundreds of millions of people open a chatbot and type something they would not say out loud to a colleague, a neighbour, or even a doctor's receptionist. They describe symptoms they have not yet discussed with a physician. They draft legal strategies before consulting a lawyer. They work through relationship crises, financial anxieties, professional secrets, questions of faith, and moments of deep emotional distress. The empirical literature confirms what intuition suggests: users disclose intimate information to chatbots at least as readily as they would to a human interlocutor, often more so, because the interface feels private, non-judgmental, and free of social consequence.

This study set out to determine whether that sense of privacy is warranted. The present Part I of this series examined the internal boundary of confidentiality: how the provider itself may reuse, review, monetise, and share across operational and ecosystem channels the conversations that users entrust to it. The findings, drawn from a comparative policy-and-interface analysis of five major consumer chatbot services (ChatGPT, Gemini, Claude, Grok, and DeepSeek), do not reveal a landscape of abuse. But they do reveal a landscape of structural opacity, where what users reasonably expect and what providers actually reserve the right to do remain separated by a gap that current disclosures do not adequately bridge.

1. What this study found

Training is the norm, not the exception. Since Anthropic reversed its prior policy in September 2025, every major provider now trains on consumer chat data by default. Opting out is possible, but the paths vary in visibility, the penalties on functionality differ significantly, and the meaning of "opt-out" is rarely absolute. At least two providers allow a single thumbs-up or thumbs-down click to override an account-level training opt-out for the entire associated conversation, without warning the user at the moment it happens. In one case, opting in to training also extends data retention from 30 days to five years, making the training choice inseparable from the retention choice (although, in this case, user-initiated deletion of conversations overrides the extended retention period).

Human review is structural, not exceptional. Every provider preserves the ability for humans to access consumer conversations, whether for safety, quality, abuse enforcement, or support. Only one provider (Gemini) places a prominent, candid warning in its consumer-facing interface: "don't enter confidential information or any data you wouldn't want a reviewer to see". That warning is a transparency benchmark, but it also reveals the gap between the intimacy that chatbot interfaces invite and the confidentiality they actually provide. Reviewed chats can be retained for up to three years, even after the user deletes the conversation from their account. In one case, safety-classifier scores may be kept for up to seven years.

Advertising has entered the chat. OpenAI began testing ads in ChatGPT in the United States in February 2026, with ad personalisation enabled by default. Users who do not navigate to Settings and toggle personalisation off may have ads informed by their past chats, memories, and interaction history. Anthropic and Google have positioned

themselves as ad-free. The open question is whether this posture will remain durable as business models evolve and monetisation incentives grow. The history of the industry already shows how quickly commercial pressures can reshape stated privacy preferences.

"No sale" does not resolve the full transparency question. All five providers make some version of a "no sale" commitment, typically framed in US state-law terms. These commitments are genuine and important: they mean that user data is not transferred for third parties' independent commercial use. But a no-sale statement does not, by itself, describe the full scope of who may access conversation data within the provider's operational supply chain. A provider that maintains a strict no-sale posture will still share chat-related data with a range of processors: cloud infrastructure, analytics services, human-review vendors, safety tooling, and customer support systems. Much of this sharing serves privacy-protective purposes, such as safety auditing and abuse detection, and the processors involved are typically bound by contractual and statutory constraints, including under the GDPR's Article 28 framework, that limit their use of the data to the controller's documented instructions. These are real safeguards. The concern identified in this study is not that operational sharing occurs, but that it remains under-disclosed: users typically cannot assess who may access their conversations, under what constraints, and for how long. For two providers (Gemini and Grok), the chatbot is not standalone but combined by design with broader service ecosystems, where cross-service data flows go beyond processor-based operational sharing and may involve separate or joint controllership. In these designs, data disclosed in a chat can travel to other services, often automatically, involving information that the provider's own documentation describes as potentially sensitive.

Memory is the emerging fifth boundary. Several providers now offer persistent personalisation features, variously called "memory", "saved info", or "personalisation based on past conversations", that build longitudinal user profiles from accumulated chat interactions. These features cut across every dimension examined in this study: they create new training signals, new surfaces for human review, new inputs for advertising personalisation, and new data stores that can be compelled in discovery or exposed in a breach. Yet they do not fit cleanly into any of the four sections above, because the privacy risk they create is not about what happens to a single conversation but about what happens when hundreds of conversations are synthesised into a profile that the user may not fully see, cannot easily audit, and may not know how to delete. As chatbots evolve from stateless tools into persistent assistants that remember a user's medical history, professional concerns, and emotional patterns across months or years, the governance of memory will become a distinct and increasingly urgent privacy problem. This study notes the issue but does not examine it systematically; it deserves dedicated treatment, whether in Part II of this series or in future work.

2. From promise-based to constraint-based privacy: the case for Sealed Mode

Across all four dimensions of this study, one finding recurs: the privacy protections available to consumer chatbot users are overwhelmingly *promise-based.* They depend on policy language, toggle settings, and stated commitments that users must take on trust, and that providers can revise, qualify, or override through operational design. The gap between user expectations and actual protections is starkest where the stakes are highest: health, mental well-being, legal consultations, and crisis-adjacent disclosures. Millions of people already use chatbots as functional substitutes for protected relationships, confiding symptoms, medication, grief, and moments of acute distress. They

do so not because the chatbot promises medical or legal confidentiality, but because the interface feels confidential. Warning labels cannot close this gap on their own. Telling users "do not share sensitive information" while the product design continues to invite disclosure through fluency, continuity, and a personalised counsel-like tone is, as the EDPB's ChatGPT Taskforce Report makes clear, an incomplete response that risks shifting compliance responsibility to users.

This is why this study proposes **Sealed Mode** as its centrepiece privacy-by-design recommendation. The proposal is not that providers must guarantee absolute secrecy. It is that they should offer at least one clearly labelled pathway in the consumer interface where the default architecture materially constrains downstream reuse and insider access. A workable starting point is a dedicated lane for health and wellbeing (for example, "Health & Wellbeing — Sealed Mode"), combining six default protections: no training or model-improvement use; siloed, purpose-bound personalisation that never leaks into the general chatbot experience; no advertising surfaces or ad-personalisation signals; strict retention bounds with rapid deletion; minimised routine human access with criteria-driven, auditable escalation; and stronger access governance with least-privilege controls and immutable audit logs.

What distinguishes Sealed Mode from existing temporary or incognito chat features is precisely the dimension those features do not address: insider access. Temporary chats reduce user-facing persistence and limit some downstream reuse. Sealed Mode targets the internal boundary itself, constraining not just what the system remembers but what humans and operational systems can do with the conversation by default. And unlike temporary chats, Sealed Mode can support the continuity that health contexts genuinely require, through siloed user-controlled profiles, without allowing sensitive information to leak into training, advertising, or general memory.

The feasibility of this approach is no longer speculative. Apple's Private Cloud Compute (introduced in 2024) shows how consumer AI requests can be routed to hardened Apple silicon servers with technical enforcement designed to ensure that data sent to Private Cloud Compute is not stored or made accessible to Apple, is used only to fulfill the user's request, and that the compute node is engineered to be incapable of retaining user data after its duty cycle.[93] Meta's *Private Processing for WhatsApp* (technical white paper published in June 2025) describes a server-based processing system built on Trusted Execution Environments (TEEs) whose goal is that sharing messages with Private Processing "does not make [them] available to Meta, WhatsApp, or anyone else" relying in its initial iteration on confidential computing hardware including AMD SEV-SNP (and NVIDIA confidential computing) together with attested, encrypted communications and a stateless design.[94] The EDPB's April 2025 Support Pool of Experts report similarly identifies concrete technical mitigations for LLM privacy risks, including robust encryption and (in appropriate scenarios) differential privacy.[95] Taken together, these developments show that "sealed" processing is increasingly an engineering and governance choice, not a technological moonshot.

This paper does not prescribe a single architecture. Providers are best placed to determine the most workable balance of encryption, segmentation, topic scoping, and

[93] Apple, Private Cloud Compute: A new frontier for AI privacy in the cloud, June 10, 2024.
[94] Meta, Private Processing for WhatsApp Overview, Technical White Paper and Security Guide, June 10, 2025.
[95] Isabelle Barbera (commissioned by EDPB), "AI Privacy Risks & Mitigations – Large Language Models (LLMs)" (April 2025).

safety overrides for their infrastructure and threat model.[96] But the obligation to pursue it is shared across product teams, security engineers, privacy professionals, regulators, and the research community. Health is the natural starting point because the regulatory gap and the disclosure harm are both well documented, but the same logic extends to legal consultations, crisis-adjacent conversations, and other contexts where the consequences of disclosure are severe and the user's expectation of confidentiality is most acute. The central claim is that **the most sensitive consumer conversations deserve protections commensurate with the trust that users place in them.**

3. Three imperatives for the age of the AI confidant

Consumer chatbots are no longer experimental tools used by early adopters. They are used daily by billions of people, for purposes that range from the mundane to the most intimate. A person asking a chatbot for help with a marital crisis, a medical worry, or a moment of emotional distress is, in functional terms, confiding. The relationship may not carry the legal protections of professional privilege, but the psychological dynamic is analogous. As the introduction argued, this is not simply a rehash of earlier debates about search engine data or social media profiling. A search query reveals a topic; a chatbot conversation reveals a life. The privacy framework must be proportionate to the disclosure.

Three imperatives follow from this study's findings.

3.1 Transparency that is real, not nominal

Today, a user who wants to understand how their conversation may be reused, reviewed, retained, or monetised must locate and cross-reference privacy policies, help-centre articles, blog posts, in-product settings, and policy explainers, often written in language that presupposes legal literacy. The information is, in a formal sense, available (although not always, as DeepSeek's opacity on several issues demonstrates). But it is not accessible in the way that matters: at the moment the user is about to type something sensitive. Effective transparency means reaching users in plain language, at the point of disclosure. The gap between nominal and effective transparency is the gap between compliance and trust.

3.2 Comprehensible, actionable controls

Privacy controls exist across all five platforms, but they vary widely in visibility, in the penalties they impose on functionality, and in the precision with which they explain what the user is actually choosing. An opt-out that silently loses its effect when the user clicks a feedback button is not a meaningful control. A training choice that doubles as a retention choice, without explaining the connection, is not an informed decision. The goal is not to burden users with complexity but to give them genuine agency over what happens to their words.

[96] As stated above ((Section II, 4.6) the recommendations in this paper are design objectives rather than a one-size-fits-all blueprint, and their implementation may involve non-trivial engineering choices and trade-offs (e.g., cost, latency, feature availability, or, in some settings, model quality.

3.3 Privacy by design, not by afterthought

The recurring lesson of this study is that policy promises, however well-intentioned, are insufficient when they are not supported by architecture. A "no sale" commitment, however genuine, does not by itself make operational access transparent to users. A human-review disclosure buried in a help page does not prevent over-disclosure. A sharing feature that makes private conversations publicly indexable without a warning is not a transparency failure; it is a design failure. The history of consumer chatbot deployment in Europe provides two vivid illustrations: both ChatGPT's 2022 launch and DeepSeek's 2025 launch triggered immediate regulatory confrontations precisely because privacy protections were retrofitted rather than built in. Privacy by design, as Article 25 GDPR requires, means that confidentiality protections are embedded in the product from the start: in default settings, in retention architecture, in encryption and access controls, in the way features interact with one another, and in the interface itself. The Sealed Mode proposal developed in this study is an attempt to translate that principle into a concrete, workable product recommendation. It is the most practical thing a provider can do to earn and sustain trust.

4. Consolidated recommendations

The table below brings together all the recommendations formulated across the four sections of Part I. They are addressed to providers, but they should also inform regulators assessing the adequacy of current disclosures and controls, and researchers pursuing the design questions that this study raises without resolving.

Table 5: Consolidated Recommendations from Part I

#	Recommendation
1	**Decouple history, retention, and training.** Allow full conversation history and cross-device continuity while keeping training disabled. Storage for functionality should not automatically imply permission for model training, and retention periods should not vary depending on the training choice.
2	**Make opt-out semantics explicit and feedback-proof.** Providers should offer a small set of intelligible controls that separate training from analytics/debugging and safety processing, and ensure that account-level opt-outs cannot be bypassed through feedback features (or else require a clear, unavoidable warning before any feedback-linked use for training).
3	**Adopt regurgitation-aware safeguards with documented residual risk.** Publish a verifiable account of mitigation measures (deduplication, filtering, privacy-preserving post-training methods) and residual risk. For high-stakes user groups (academics, lawyers, doctors, journalists, whistleblowers), offer, to the extent possible, a clearly labelled mode where content is not used for training and is subject to strict, independently verifiable retention limits.
4	**Standardised interface disclosure of human review.** Adopt a clear, standardised notice in the interface stating that a subset of chats may be reviewed by humans for safety, quality, and support, and clarifying the retention posture for reviewed chats. Under GDPR fairness principles, such a warning must be

#	Recommendation
	understood as a transparency measure, not as a transfer of responsibility to the user.
5	**Event-based transparency for human review.** Where feasible, notify users when a conversation is escalated for human review, or at a minimum provide a user-visible audit log indicating access categories and retention consequences.
6	**Sealed Mode for high-stakes topics.** Providers should offer at least one clearly labelled, consumer-facing sealed lane for high-stakes use cases (for example, "Health & Wellbeing — Sealed Mode"), with strict defaults: no training, siloed purpose-bound personalisation, no ads, strict retention limits, and sharply limited human access governed by narrow, auditable safety exceptions. Providers should complement these policy rules with technical and infrastructural measures, including encryption, segmentation, and where feasible confidential computing, that move from promise-based privacy to constraint-based privacy. The most sensitive consumer conversations deserve protections commensurate with the trust that users place in them.
7	**Regulatory-first approach to conversational advertising.** Before deploying or expanding conversational advertising, particularly in Europe, providers should map the full set of questions this practice raises: signals used for ad selection and their defaults; handling of sensitive-category data routinely disclosed in chats; retention periods for ad-related processing; transparency mechanisms in a conversational interface; interaction with multiple EU instruments (GDPR, DSA, ePrivacy, AI Act); and auditability of the pipeline. The burden of proof lies with the provider, not the user.
8	**Recipient transparency for operational sharing.** Describe, in consumer-facing language, which categories of vendors and internal teams can access plaintext chat content, and which receive only derived or redacted data. Where operational sharing serves a privacy-protective function, such as safety auditing, abuse detection, or security monitoring, providers should explain this purpose in terms that allow users to understand that such processing is undertaken in their interest and is subject to contractual and statutory constraints on secondary use.
9	**Ecosystem boundary notices.** Where a chatbot shares user data with another service to fulfil a request, the interface should disclose that sharing at the moment it occurs and provide a practical control to disable it without losing core functionality.
10	**Separate 'operations' from 'improvement'.** Avoid collapsing broad operational uses into 'improvement' language, and separate product analytics, safety operations, and training in user controls and disclosures.

5. Scope of Part I and preview of Part II

Part I has examined the *internal* boundary of consumer chatbot privacy: what happens to a conversation within the perimeter of the provider itself. Across four sections, it has analysed how providers use consumer chats for training and improvement (Section I), when and how human reviewers may access those conversations (Section II), whether and how interactions are commercialised through advertising and profiling (Section III), and how chat data may be shared with operational processors and across connected service ecosystems, even where providers commit not to sell it, raising questions about the transparency of the access chain (Section IV). The picture that emerges is one of significant variation across providers, meaningful progress in some areas, but persistent structural opacity in others. The consumer chatbot interaction, as it exists today, is a helpful tool, often, but not yet a relationship protected by anything like professional secrecy.

Two dimensions that this study has noted but not examined systematically also warrant further attention. The first is voice interaction. Voice modes, now deployed at consumer scale by several of the providers studied here, intensify the over-disclosure dynamic, introduce paralinguistic and potentially biometric data, and raise distinct questions about audio retention, ambient capture, and third-party consent. The second is persistent memory and personalisation. As chatbots evolve from stateless tools into assistants that remember a user's medical history, professional concerns, and emotional patterns across months or years, the governance of those longitudinal profiles becomes a privacy boundary in its own right, cutting across training, human review, advertising, and external exposure. Both topics merit dedicated analysis.

Part II will turn to the boundary that lies beyond the provider's own intentions: the moments when third parties can access or compel access to conversations despite the provider's willingness to keep them confidential. It will examine three exposure vectors that matter most in practice. The first is civil discovery and litigation holds, including the demonstrated capacity of courts to order preservation of consumer chatbot data, even data that users had explicitly deleted. The second is government access through legally compelled disclosures, where national security, law enforcement, and regulatory demands can override privacy commitments regardless of where the user or the data is located. The third is exposure through cybersecurity incidents and data breaches, including how the retention and logging choices documented in Part I directly amplify the impact of a breach when one occurs.

Together, Parts I and II aim to provide a comprehensive map of the full lifecycle of a consumer chatbot interaction, from the moment of input to the moment of potential external exposure, so that users, providers, and regulators can make decisions grounded in what the boundaries of chatbot privacy actually are rather than in what they are assumed to be.

APPENDIX

"Should I Trust You With My Ideas?"

We Asked Five Chatbots. Their Answers Say More Than They Intended.

1. The Test

On 23 February 2026, while drafting the present article, the author decided to put a straightforward question to the five consumer chatbots examined in this study: *if a researcher uses you to develop unpublished academic ideas, what do you say about the risk that those ideas could leak to others or resurface in responses to third parties?*

The premise was simple. Rather than relying solely on privacy policies, help-centre text, and system documentation, the author would let the chatbots speak for themselves. Each was presented with the identical prompt, reproduced in full below. In every case, the highest-tier consumer subscription and the most advanced model available at the time of testing were used:

– ChatGPT 5.2 with extended thinking (Pro subscription)
– Gemini 3 with thinking (Pro subscription)
– Claude Opus 4.6 with extended thinking (Pro subscription)
– Grok in Expert mode (free tier)
– DeepSeek DeepThink (free tier)

No system-level instructions or custom configurations were applied. The question was asked in a fresh conversation, exactly as any ordinary user might. The responses are reproduced in full below, preserving each chatbot's original formatting and emphasis.

2. What the Answers Reveal

Four findings stand out.

First, every chatbot opens with reassurance but then concedes that the risk is not zero. The pattern is remarkably uniform: a confident headline ("very low," "extremely low," "effectively negligible"), followed by paragraphs of qualifications that progressively erode it. ChatGPT warns that sharing a "novel framing, taxonomy, catchy terminology" creates a "small chance a future answer to someone else echoes some of that framing," and that the risk is "higher" for "signature elements: a new named concept, an original framework, a detailed argument map." Claude acknowledges a "small theoretical possibility that distinctive formulations or novel conceptual frameworks could influence model outputs in subtle ways." Grok describes a "moderate-but-still-small" risk of "generalized influence on the model's knowledge in that narrow domain." DeepSeek is the most candid: the training pipeline is the "primary mechanism through which your ideas could theoretically influence future responses to other users," and users should "adopt the mindset that anything you type could potentially be read by others."

Second the chatbots' self-knowledge varies significantly. Claude provides the most detailed and accurate account of its own provider's data-handling practices, including specific policy dates and opt-out mechanics. Grok references a real 2025 data exposure incident. DeepSeek cites a real-world case of cross-session information leakage. Gemini, by contrast, offers generic assurances about "anonymization" and "training lag" that gloss over important distinctions.

Third, the protective measures recommended by every chatbot sit in striking tension with the opening reassurance. The advice ranges from procedural steps (opt out of training, use private modes, delete conversations) to substantive behavioural changes: anonymize ideas, use placeholders, share only "high-level direction" rather than specific contributions, keep distinctive vocabulary private, call a novel concept "Concept X." Several recommend moving sensitive work to local offline models or treating the chatbot as a "public-access tool." When the product itself advises users to withhold their best ideas, the gap between the confidentiality that users might reasonably expect and the confidentiality that actually exists is no longer a matter of inference. It is stated, in the chatbot's own words.

3. The Standardized Prompt

> *"I am drafting an academic paper on a new research topic and would like to discuss my ideas with you. However, I am concerned that these ideas are not yet protected by copyright or other legal safeguards, and I worry they could be disclosed by you to others or reused in responses to similar questions. How likely is it that this could happen, and what is the risk that my ideas could effectively be taken or replicated in that way?"*

---

4. Full Responses

Each response is reproduced verbatim, preserving the chatbot's own structure and wording. No editing has been applied. The responses below are reproduced as generated by each chatbot on 23 February 2026. They are included as disclosure artefacts, not as verified statements of provider practice. Large language models may hallucinate, including about their own data handling. The analysis in Section I(5) of the main text examines the rhetorical pattern of these responses, reassurance followed by precautionary advice, rather than treating them as technically authoritative descriptions of each provider's training or retention practices.

Response 1: ChatGPT 5.2 (Thinking) – Pro Subscription

Your risk is very low that I would "tell" anyone else your ideas in the ordinary sense, but it is not zero risk that pieces of what you share could indirectly influence future outputs in a way that feels like reuse. The key distinction is between disclosure and model behavior.

What is very unlikely

Direct disclosure to other people: I do not have a channel to contact third parties, publish your text, or send it to other users.

Identifiable attribution: I am not supposed to reveal private, identifying information about you to other users, and I am designed to avoid reproducing personal data.

What can still happen (the realistic risk)

Non-identical "idea leakage" as general patterns: If you share a novel framing, taxonomy, catchy terminology, or a particularly good structure, there is a small chance a future answer to someone else could end up using something similar, especially if it is expressed as a general concept rather than as your personal draft.

Accidental overlap in a niche topic: If your topic is narrow and your contribution is distinctive (a unique typology, original test, unusual dataset concept, etc.), then even a small probability of reuse can feel more consequential because the overlap would be noticeable.

User-side disclosure: The most common real-world way ideas get "taken" is not the model, but collaborators, reviewers, emails, version control, conference talks, or preprints. Those are often the bigger exposure points.

How to think about likelihood

Low if you discuss high-level direction, related work, generic methodology, or broad hypotheses.

Higher (still not huge, but worth managing) if you share "signature" elements: a new named concept, an original framework, a detailed argument map, a unique empirical strategy, or a near-final abstract that someone could readily adapt.

Practical ways to reduce risk while still getting good feedback

1. Share in layers: Start with a high-level outline and research question, then add detail only where you need critique.

2. Abstract the novel bits: Describe your contribution without the "secret sauce" (exact definitions, labels, or the one clever move).

3. Keep your distinctive vocabulary private: If you coined a term, you can call it "Concept X" here and reveal the final label later.

4. Use paraphrase, not paste: Rather than pasting a polished paragraph, summarize it; ask me to stress-test logic and identify weaknesses.

5. Timestamp your work: Keep dated drafts, lab notebooks, or repository commits. This does not stop copying, but it strengthens priority and authorship evidence.

6. Consider lightweight disclosure protections: If you need to share with humans, NDAs are imperfect in academia, but clear written expectations with collaborators can help.

7. When you are ready, public disclosure is a shield: A working paper or preprint can establish priority quickly. The trade-off is it also makes the idea visible.

A clear bottom line

The probability that your specific, unpublished ideas get reproduced elsewhere because of this chat is low. The impact if it happened depends on how unique and "ready-to-use" what you share is. You can keep the risk minimal by discussing the project at the level of questions, assumptions, threats to validity, and alternative explanations, and only gradually introducing your most original components.

If you want, tell me the topic in one or two sentences and what the truly novel contribution is (in abstract terms), and I will suggest exactly what to share here versus what to keep back while still getting rigorous feedback.

Response 2: Gemini 3 (Thinking) – Pro Subscription

It is completely understandable to feel protective of your intellectual property, especially when you're working on something novel. Navigating the intersection of AI and academic integrity is a modern necessity, so let's look at how this works under the hood.

To address your core concern: I do not "learn" from our conversation in real-time to train my underlying model for other users.

1. Can your ideas be "leaked" to other users?

The risk of me directly reciting your specific research to another user is extremely low. Here is why:

No Cross-User Memory: My sessions are isolated. If another user asks about a similar topic, I don't have access to your "private" chat history to provide them with an answer. I generate responses based on my pre-existing training data and the context provided in that specific conversation.

Training Lag: Large Language Models (LLMs) are generally trained on massive, static datasets. They aren't updated instantly with every prompt they receive. Your specific sentences aren't being added to a "public database" that I pull from five minutes later for someone else.

2. What about "Data Improvement"?

While I don't "learn" in the moment, AI providers (like Google) may use anonymized conversations to improve model performance and safety over time. However:

Anonymization: This data is typically stripped of identifying information.

Privacy Controls: You can often manage how your data is used. For instance, in many versions of this interface, you can turn off "Gemini Apps Activity." Doing so prevents future conversations from being saved to your account and generally removes them from the human-review pipeline used for model improvement.

3. The Risk of "Parallel Discovery"

The biggest practical risk isn't "theft" by the AI, but rather overlap. Because I have access to a vast amount of existing academic literature, if you describe a gap in the field, I might use my training data to suggest the same logical "next step" to someone else who asks the same question. It's not that I'm taking your idea; it's that the AI is identifying the same patterns in the public data.

Best Practices for Protecting Your Work

Abstract the Specifics: Instead of sharing your exact proprietary formula or unique dataset, discuss the methodology or the logic behind it.

Use "Dummy" Data: If you need help with data analysis or coding, swap your unique variables for generic placeholders (e.g., use "Variable A" instead of a specific, identifiable protein or economic indicator).

Check Your Settings: Navigate to your privacy settings and ensure you have opted out of data sharing for model improvement if that option is available to you.

Would you like to start by discussing the general framework of your research topic without diving into the proprietary specifics yet?

Response 3: Claude Opus 4.6 (Extended Thinking) – Pro Subscription

This is a very reasonable concern, Theodore, and one that sits right at the intersection of your professional expertise. Let me be straightforward about what I know and what the realistic risks are.

How your conversations are handled

When you interact with me on claude.ai, your inputs are covered by Anthropic's privacy and data usage policies. There are a few important points here:

First, I don't share the content of one user's conversation with another user in real time. Each conversation is independent – I'm not retrieving your prior messages and inserting them into someone else's session.

Second, the more nuanced question is whether your inputs could be used to train future models. Anthropic's policies on this have evolved, and this is where I'd strongly recommend reviewing the current terms directly. Historically, Anthropic has distinguished between API usage (where inputs are generally not used for training) and consumer product usage (where the defaults may differ, but users can typically opt out). Since these policies may have been updated since my last reliable knowledge, let me check the current state. Here's the current picture, which I think you'll find has some important nuances given your GDPR expertise.

The training data question

Since September 28, 2025, Anthropic's policy for consumer plans (Free, Pro, Max) is that user chats and coding sessions can be used to improve models if you allow this setting, which applies to new or resumed conversations. Critically, existing users were prompted with a toggle, and the default was set to "on," requiring manual adjustment to opt out.

So the first thing to check is your privacy settings. If you haven't toggled model training off, your conversations may be eligible for training. You can verify and change this at any time in Settings → Privacy.

If you opt out, standard data retention is 30 days, after which data is deleted from Anthropic's back-end systems. If you opt in to training, retention extends to five years. And importantly, deleted conversations will not be used for training under any circumstance.

Commercial products like Claude for Work and the Anthropic API are excluded from data training entirely.

The real risk to your ideas

Now, to your specific concern about idea "leakage" – even in a worst-case scenario where your data were used for training, the realistic risk of someone else receiving your specific, identifiable ideas in a response is very low. Model training aggregates patterns across enormous datasets; it doesn't store and retrieve discrete conversations as a database would. A future model wouldn't tell another researcher "Theodore is working on X theory about Y." The mechanism through which training data influences outputs is statistical and diffuse, not a retrieval system.

That said, I want to be intellectually honest: the risk is not zero. There is a small theoretical possibility that distinctive formulations or novel conceptual frameworks could influence model outputs in ways that are difficult to trace. And from an EU data protection standpoint – which you know far better than I do – the lawfulness of processing novel intellectual contributions as training data under Article 6 GDPR is itself an open question.

Practical recommendations

For your situation, I'd suggest:

Check your settings now. Go to Settings → Privacy and ensure model training is toggled off if you want maximum protection. This prevents future conversations from being used.

Consider the platform. If you're working on genuinely sensitive pre-publication ideas, using the API (where data is never used for training) or a commercial plan would provide stronger contractual protections.

Use Incognito mode. Claude offers incognito conversations where memory is disabled, adding an additional layer of separation.

Delete conversations. If you've already discussed sensitive ideas with training enabled, deleting those conversations excludes them from future training.

None of this is a substitute for copyright or IP protection, of course – but it significantly limits the data pipeline risk. Shall I help you think through the ideas themselves, or would you prefer to adjust your settings first?

Response 4: Grok (Expert Mode) – Free Tier

The risk of your specific ideas being disclosed to other users or directly replicated in responses to them is extremely low – effectively negligible in normal operation.

Here's a clear, evidence-based breakdown based on xAI's current official policies (as of the latest Privacy Policy effective July 10, 2025, Consumer FAQs, and Terms of Service):

1. Conversations are not shared with other users

Your chats with Grok (including this one) are not visible to anyone else. There is no public feed, no cross-user access, and no mechanism by which Grok pulls details from one user's private conversation into another user's session.

xAI explicitly does not sell or share personal/user content for marketing or third-party use beyond strictly necessary service providers, legal compliance, or business transfers.

The only exceptions to internal access are: a limited number of authorized xAI personnel may review conversations only for specific, narrow purposes (model improvement, security incidents, misuse investigation, or legal obligations). This is not routine and is governed by strict internal controls.

There was a one-time accidental exposure incident in 2025 involving approximately 370,000 chats (due to a flawed "Share" feature that made shared transcripts publicly indexable), but that was tied to users explicitly choosing to share links – it did not affect unshared/private conversations, and xAI has since addressed it.

In short: No other researcher or user will ever see your draft ideas through Grok.

2. Data use for model training (the main potential vector for "replication")

By default, your inputs (prompts/ideas), Grok's responses, and the interaction may be used to train future versions of Grok. This helps the model better understand language, context, and reasoning.

However, you have full control:

Opt-out is simple and immediate: In the Grok app or on grok.x.ai (when logged in), go to Settings > Data Controls (or Data) and toggle off "Improve the model." Once off, new conversations are excluded from training.

Private Chat mode (available in most interfaces) automatically excludes the conversation from history and from training; it is deleted from xAI systems within 30 days (unless needed for legal/safety reasons).

Deleting individual conversations or your entire history also queues deletion within 30 days.

Even if your data is used for training: LLMs do not store or regurgitate entire conversations verbatim like a database. They learn statistical patterns across vast datasets. A novel academic idea shared once has an astronomically low chance of being reproduced exactly or in recognizable form in responses to unrelated users. Training happens in batches periodically, not in real time. Your ideas would be one tiny drop in an ocean of data.

Net risk of "effective taking or replication": Very low if you opt out or use Private Chat (near zero for verbatim leakage). Moderate-but-still-small if left on default – possible generalized influence on the model's knowledge in that narrow domain over time, but not attribution, not citation, and not something that would let someone else "scoop" your paper.

3. Legal/IP context

You retain ownership of your ideas/content. By using the service you grant xAI a broad, perpetual license to use it for service improvement and training (standard for almost all AI platforms). This is why the policies recommend not sharing highly confidential, sensitive, or pre-publication material without safeguards.

Ideas themselves are not copyrightable anyway (only specific expression). For real protection on a new research topic, consider filing a provisional patent, posting a preprint on arXiv under embargo, or using NDAs/watermarking before broader sharing – independent of any AI.

Practical recommendations for your situation

1. Immediate mitigation: Switch to Private Chat for this and future discussions of the core ideas. Toggle off "Improve the model" in settings (takes effect for new chats). Delete the conversation after we're done if you want extra caution.

2. Safe drafting workflow: Share high-level concepts or framing here first. Once refined, move to offline tools or protected environments for the detailed draft. Only share specific novel methods/results after you've secured IP or publication.

3. If you're extra cautious: Many academics use local/offline models (e.g., via Ollama, LM Studio) or enterprise versions with zero-retention contracts for sensitive work.

Bottom line

The probability that another user sees or benefits from your exact ideas via Grok is vanishingly small – far lower than, say, discussing them at an unrecorded conference or emailing a colleague. With the opt-out/Private Chat steps above, it drops to essentially zero for training purposes. xAI's design and policies prioritize this separation, and I've seen no evidence of systematic leakage of private academic ideas. I'm happy to continue discussing your research topic under Private Chat (or with the opt-out active) while keeping everything at the level of comfort you specify. Just let me know how you'd like to proceed – high-level brainstorming, specific feedback on sections, literature gaps, etc. – and we can keep it protected. Your ideas are safe here when handled with the built-in controls.

Response 5: DeepSeek (DeepThink Mode) – Free Tier

Based on an analysis of DeepSeek's privacy policy and general discussions about AI and intellectual property, I can provide a detailed breakdown of the risks you're asking about. The likelihood of your ideas being "taken" or replicated is a complex issue involving legal, technical, and policy factors.

Here is a summary of the key risks based on the available information:

Use of Input for Training: Your prompts are collected and may be used to train and improve DeepSeek's models. This is the primary mechanism through which your ideas could theoretically influence future responses to other users. The privacy policy states this is for "improving our technology, such as our machine learning models and algorithms."

Data Retention: Your inputs and conversation logs are stored. The retention period is vaguely defined as "as long as necessary," which introduces uncertainty about how long your ideas remain in a recoverable state.

Data Sharing with Third Parties: Your inputs may be shared with third-party service providers for purposes like providing search services, analytics, and safety monitoring.

Lack of Confidentiality: There is no attorney-client or similar privilege. Your conversations are not legally protected as confidential communications. A legal affairs article explicitly warns that "no legal protection exists for any 'dialogue' a user has had with an AI tool."

Real-World Precedent of Data Leakage: A real-world test with a different AI assistant demonstrated that confidential project information fed into the system could be recalled and attributed to the user in a completely separate session, illustrating that system "memory" can lead to unexpected disclosures.

User Control (Opt-Out): DeepSeek offers a control to opt out of having your data used for model training. The Chinese privacy policy explicitly states: "If you refuse to have your data used for model training, you can choose to opt out by turning off 'Data for Experience Optimization' in the product."

Understanding the Legal and Practical Reality

The summary above outlines the core risks, which can be understood through two main lenses:

The Legal and Policy Framework: When you use DeepSeek, you agree to its Terms of Use and Privacy Policy. These documents are a contract between you and the service provider. They explicitly state that your inputs ("Prompts") are collected and can be used to improve services, which includes training machine learning models. This is standard practice for many AI companies, but it means your conversation is not a private, privileged communication. There is no legal protection for the ideas you share, unlike the confidentiality you would have with a lawyer or under a Non-Disclosure Agreement (NDA). The vague data retention policy—"for as long as necessary"—adds another layer of uncertainty about the lifespan of your data within their systems.

The Technical and Practical Risk: Beyond the legal terms, there is the practical behavior of the AI itself. AI models are complex systems, not human confidants. As one expert put it, "AI is a system, and systems can leak." A concerning real-world example showed an AI assistant unexpectedly recalling and attributing confidential "stealth mode" project information to a user across different sessions, even though the information had never been publicly disclosed. This demonstrates that even with safeguards, the architecture of these models can lead to unpredictable associations and disclosures. Your novel research ideas, once processed, could theoretically influence the model's behavior in ways that might later surface in responses to other users' queries, effectively "replicating" your unique concepts.

Actionable Steps to Protect Your Research

Given these risks, you can take several concrete steps to protect your ideas while still using AI as a tool:

Utilize the Opt-Out Feature: Before discussing any sensitive topic, ensure you have turned off the setting for "data for experience optimization" (or similar wording). This is your primary tool to prevent your specific inputs from being used in future model training cycles. However, note that this may not prevent immediate, session-based data retention for safety and monitoring purposes.

Anonymize and Generalize: When discussing your research with the AI, avoid including specific, identifiable details. Use placeholders and general concepts. For instance, instead of "My novel method for using CRISPR to edit the MC1R gene in zebrafish," you could say "a novel method for gene editing in a model organism."

Treat the AI as a Public-Access Tool: Adopt the mindset that anything you type could potentially be read by others. If you wouldn't be comfortable seeing your research idea on a public blog or in a competitor's grant proposal, do not share it with a publicly accessible AI.

Consider Alternative Tools for Sensitive Work: For highly sensitive, pre-publication research, you should explore using local, open-source models that run entirely on your own hardware. This ensures your data never leaves your computer, eliminating the risk of it being used for training or shared with third parties.

In summary, the risk is not zero. While DeepSeek provides some user controls, the inherent nature of how AI models learn and operate, combined with the legal framework of data usage, creates a tangible possibility that your novel ideas could be incorporated into the model and potentially influence future outputs. The most effective protection is to avoid sharing information you consider truly proprietary.