ARTIFACT EVALUATED
USENIX
AVAILABLE

What Users Ask, Policies Miss: Unveiling the Gap Between Community-Expressed Privacy Concerns and LLM Provider Policies

Zhihuang Liu[†], Zhen Huang[†], Ling Hu, Yifan Yang, Zhiping Cai[*]
National University of Defense Technology
{lzhliu,huangzhen25,linghu50,yangyifanyyf,zpcai}@nudt.edu.cn

Abstract

Large Language Models (LLMs) process millions of conversations containing sensitive information daily, yet whether their privacy policies adequately address users' publicly expressed concerns remains unexplored. This paper presents the first large-scale, user-centered audit of privacy policy adequacy in LLM services. We systematically extract privacy concerns from Reddit communities of five major LLM providers and assess whether their latest privacy policies address these concerns. Our semi-automated pipeline analyzes 1,531 threads to extract 4,994 authentic privacy concerns, which we organize into a 20-topic taxonomy spanning four thematic groups. We then identify 3,137 policy gap instances and classify them into six categories: four policy coverage gaps (detail vague, AI feature unaddressed, vulnerable group neglected, and jurisdiction unclear) and two user perception gaps (explicit distrust and awareness deficit). Our analysis reveals that coverage gaps and perception gaps contribute nearly equally (50.2% vs. 49.8%), with AI-specific feature gaps (36.6%) and user awareness deficits (40.7%) together comprising the vast majority (77.3%) of all identified gaps. Notably, we also surface user-reported evidence suggesting potential discrepancies between stated policies and observed system behavior, highlighting the need for verifiable privacy guarantees. These findings demonstrate that improving LLM privacy requires dual-pronged interventions addressing both inadequate policy disclosures and user comprehension barriers, offering actionable insights for relevant stakeholders.

1 Introduction

Large Language Models (LLMs) and their applications, such as ChatGPT [3], Claude [5], and Gemini [7], have rapidly become integral to both professional workflows and personal digital experiences [10, 15, 72], offering users unprecedented convenience in content creation, decision support, and information retrieval [22, 49]. Unlike traditional software services, users interact with LLMs through open-ended conversational exchanges, where they may inadvertently disclose sensitive information ranging from health conditions [50] to workplace confidential matters [11]. As LLM capabilities expand into multimodal processing [65, 70] and autonomous agent functionalities [35, 47, 69], the scope of information these systems collect continues to broaden. Therefore, LLM privacy policies become essential documents through which users seek to understand providers' data practices and make informed decisions about service adoption [66].

However, a series of privacy incidents indicates that existing policy disclosures may not adequately address regulatory and user needs, underscoring the urgency of LLM privacy issues. On the regulatory side, Italy's data protection authority fined OpenAI €15 million for insufficient disclosure of data processing practices, citing violations of General Data Protection Regulation (GDPR [6]) transparency obligations [56]. On the user side, a March 2023 bug in ChatGPT exposed users' chat history titles to other users due to a flaw in an open-source library, prompting OpenAI to temporarily shut down the service [9]. Subsequently, Samsung banned employee use of ChatGPT after discovering that staff had uploaded sensitive source code to the platform [8].

In response, researchers have begun systematically analyzing LLM privacy policies, focusing on transparency and user comprehension issues. Specifically, prior work reveals that LLM privacy policies are typically lengthy, difficult to read, and filled with vague language [66]. Critical information is often scattered across multiple sub-policies and FAQ documents [36]. Moreover, disclosures regarding LLM-specific aspects such as training data usage and retention periods remain insufficient [74]. Consequently, users struggle to fully understand policy descriptions and may overlook AI-specific risks [36, 66, 74].

Meanwhile, existing studies have examined user privacy concerns about LLMs through surveys and semi-structured interviews [18, 41, 49, 75]. Although these controlled investiga-

[†]Equal contribution.
[*]Corresponding author.

tions provide valuable insights, they capture user perceptions at single points in time with limited participant diversity and scale [58]. To address these limitations, Ali et al. [11] examined user privacy concerns expressed in social media posts from `r/ChatGPT`, capturing a more diverse user base and enabling a comprehensive understanding of users' concerns, behaviors, and preferences.

Nevertheless, these two strands remain disconnected: policy studies evaluate disclosures without grounding them in community-expressed concerns, while concern studies characterize user worries without examining whether provider-specific policies address them. Consequently, a fundamental question remains unanswered: ***Do LLM privacy policies adequately address what users genuinely care about?*** Research has not yet attempted to reveal whether providers' latest privacy policies sufficiently cover user concerns as continuously expressed in public communities. Yet this question matters because privacy policies should not merely satisfy regulatory checklists but genuinely help users understand and trust how their data is handled. Understanding whether policies address what users actually worry about is therefore critical for establishing trust in LLM services. At this pivotal moment in the rapid evolution of LLM applications, **a systematic understanding of policy-concern gaps can guide the path toward more transparent and user-centered privacy practices.**

To this end, this paper presents the first large-scale, user-centered audit of privacy policy adequacy in LLM services. We systematically extract privacy concerns widely expressed by users from seven Reddit communities spanning five major LLM providers (ChatGPT, Claude, Gemini, Grok, and DeepSeek) and assess whether their latest privacy policies adequately address these concerns. Our investigation is guided by two research questions (RQs).

- **RQ1 (Community-Expressed Concerns):** What privacy concerns do users authentically express across different LLM provider communities? How do concern patterns vary across providers, and which AI-specific risks emerge as particularly prevalent?
- **RQ2 (Policy-Concern Gaps):** To what extent do LLM providers' privacy policies address user-expressed concerns? What distinguishes policy coverage gaps and user awareness barriers, and which concern topics exhibit systematic neglect?

To address the above RQs, this paper makes the following contributions.

- Conducts the first large-scale, community-driven audit that links user-expressed LLM privacy concerns to provider-specific policy evidence. Analyzing Reddit communities of five major LLM providers across 1,531 threads, we extract 4,994 privacy concerns and identify 3,137 policy gap instances, revealing systematic disconnects between user expectations and provider disclosures.
- Develops two reusable taxonomies through iterative qualitative analysis: a 20-topic privacy concern taxonomy organized into four thematic groups and a 6-type privacy policy gap taxonomy that distinguishes policy coverage deficiencies from user perception barriers.
- Reveals that policy-side and user-side problems contribute nearly equally (50.2% vs. 49.8%), with AI-specific feature gaps (36.6%) and user awareness deficits (40.7%) comprising the vast majority of all gaps, while surfacing user-reported evidence of potential policy-behavior discrepancies. Our findings provide empirically grounded insights for providers, users, regulators, and privacy researchers.

2 Background and Related Work

LLM Privacy Policy Analysis. Privacy policies serve as the primary mechanism through which service providers disclose their data practices, functioning as both a legal instrument and a communication channel between providers and users. Given the sensitive nature of conversational data and AI-specific risks (e.g., training data usage, model memorization), LLM privacy policies warrant particular scrutiny.

Characteristics of LLM Privacy Policies. Recent studies reveal that LLM privacy policies exhibit distinct characteristics compared to traditional software services. Zhang et al. [74] identify key differences in how LLM policies address input data handling, privacy protections, and model training practices. King et al. [36] found that all major LLM developers default to using user chat data for model training, with some retaining this data indefinitely. They further reveal that critical information is often scattered across multiple sub-policies and FAQ documents. Tao et al. [66] conduct the first longitudinal measurement across 11 LLM providers, finding that these policies are significantly longer, harder to read, and more ambiguous (approximately 75% of sentences contain vaguer terms) than traditional services. They also observe that providers tend to reduce explicit regulatory references, shifting toward broader phrases like "applicable law."

Policy Collection and Analysis Methodology. Prior works primarily adopt manual exploration to collect LLM privacy policies due to their structural inconsistency and distributed hosting [26, 59]. For analysis, methodologies have evolved from manual coding [36] to automated NLP-based approaches [14, 16, 20, 30], enabling scalable assessment of readability and compliance. More recently, LLM-based approaches have demonstrated potential for nuanced policy understanding [71], though no work has yet applied LLMs to analyze LLM privacy policies themselves.

Complementing policy analysis, user-centered studies of LLM privacy concerns help ground policy evaluation in users' actual perceptions and expectations rather than solely regulatory requirements.

User Attitudes and Privacy Concerns regarding LLMs. Existing efforts [18, 41, 50, 75] explore users' experiences, perceptions, and expectations regarding LLM privacy from

multiple perspectives. Specifically, some studies focus on revealing users' understanding of and attitudes toward LLM privacy policies. Zhang et al. [74] conduct the first user study examining how users interact with LLM privacy policies, revealing that users often engage superficially and miss critical information. Even after careful reading, core concerns regarding data training practices and personal data deletion persist. Chen et al. [17] propose the CLEAR tool, which detects sensitive information in user inputs and provides real-time privacy risk alerts based on relevant policy fragments. Beyond controlled surveys, researchers increasingly turn to social media for naturalistic expressions of user concerns. Ali et al. [11] analyze 2.5 million Reddit posts from `r/ChatGPT`, revealing evolving patterns in user worries with emphasis on data collection, model training, and transparency. Their work is similar to ours in its use of Reddit and LLM-assisted concern analysis. However, **our research question, scope, method, and findings differ**: we study policy adequacy gaps rather than concern patterns alone, cover 5 providers and 33 policy documents rather than a single `r/ChatGPT` community, use a multi-step pipeline that links concern extraction to necessity assessment, policy review, gap classification, and validation, and produce both concern and gap taxonomies with cross-provider gap insights.

Reddit as a Research Platform of User Perspectives. Following established ethical frameworks for publicly available data [25, 27], prior work widely adopts Reddit for observing authentic user discourse [28, 37, 45, 64]. Its topical organization through subreddits (e.g., `r/ChatGPT`, `r/ClaudeAI`) enables focused collection of domain-specific discussions. Reddit organizes discussions as *threaded* conversations: **each *thread* consists of an initial *post* and associated *comments* over time. This paper uses the term *record* to refer to either a thread or a comment.**

While prior research has studied user privacy concerns and analyzed privacy policies as *separate* endeavors, **systematic research *connecting* these two strands remains scarce**. No comprehensive study has audited whether LLM privacy policies adequately address the concerns users express in online communities. A meta-analysis of existing studies cannot answer this alignment question because prior work uses heterogeneous time periods, provider scopes, data sources, and coding schemes. Moreover, LLM policies change frequently and are fragmented across provider documents, whereas gap detection requires contemporaneous mapping from community-expressed user concerns to provider-specific policy text. Literature aggregation also cannot distinguish policy omissions or vague disclosures from user awareness deficits or distrust.

LLM-Assisted Methods for Security and Privacy Research. Literature analyzing free-text corpus like Reddit posts or policy datasets [26, 33, 34, 50, 52, 60] typically relies on manual qualitative analysis through human thematic coding (codebook development). However, manual inductive coding often suffers from scalability limitations—researchers typically sample limited subsets (e.g., [64] selects 115 posts from 3,321; [45] examines 180 posts from 7,255). Automated NLP-based methods thus have been explored for both privacy policies [14, 16, 20] and Reddit data [40, 43]. More recently, text-comprehension tasks in security research are increasingly delegated to LLMs [19, 21, 63]. LLMs are being integrated into security and privacy research workflows to assist human analysts [23, 46, 48, 55, 61, 71]. For instance, Liu et al. [48] demonstrate ChatGPT's potential in vulnerability management, while Singh et al. [61] reveal how LLMs are increasingly integrated as cognitive assistants in Security Operations Centre workflows.

LLM-Assisted Analysis for Privacy Policies and Reddit Data. More directly relevant to our work, researchers have begun applying LLMs to the analysis of privacy policies and public Reddit discussions. Xie et al. [71] developed an LLM-based approach to assess privacy policy clauses across over 100,000 websites, demonstrating more accurate analysis than rule-based methods. Nagaraj Rao et al. [53] proposed QuaLLM, an LLM-based framework applied to over one million Reddit comments. Ali et al. [11] employed GPT-4o to analyze posts from `r/ChatGPT` on Reddit, while Chen et al. [17] utilized LLM few-shot learning techniques to extract relevant segments from full privacy policies based on sensitive information identified in user inputs.

As an approach that enhances large-scale analysis in this paper, LLMs are applied to explore users' privacy concerns within massive comment data. Furthermore, we analyze lengthy privacy policies by leveraging LLMs to determine whether they fail to address these identified concerns.

3 Methodology

Our methodology comprises two main phases (*refer to our repository for an overview diagram*). First, we collect user discussions from Reddit communities of five major LLM providers and gather their corresponding privacy policy documents (Section 3.1.1-3.1.2). Second, we develop an LLM-assisted semi-automated pipeline to extract privacy concerns from Reddit threads, classify them into a 20-topic taxonomy, detect policy coverage gaps against provider policies, and categorize gaps into a 6-type taxonomy (Section 3.2.1-3.2.4).

3.1 Data Collection

3.1.1 Reddit Thread Collection

This paper selects Reddit as the data source (refer to Section 2 for specific reasons and concepts) to collect user discussions about privacy concerns related to LLMs and their providers.

Subreddit Selection. We select subreddits using the following inclusion criteria: (1) directly related to mainstream LLM products, (2) focused on discussions about specific LLMs, (3) having substantial membership with active discussions on

relevant topics, and (4) corresponding to LLM providers with publicly available privacy policies. We begin from a broader pool of approximately 10 mainstream LLM service providers and their associated Reddit communities, then apply these criteria to select provider-subreddit pairs. This process yields 7 subreddits: `r/ChatGPT`, `r/OpenAI`, `r/GeminiAI`, `r/Bard`, `r/ClaudeAI`, `r/grok`, and `r/DeepSeek`, collectively covering 5 LLM products. Other subreddits are excluded due to weak relevance to specific LLM discussions or insufficient activity. Thus, the study analyzes a criterion-based subset of mainstream, policy-covered LLM services with active Reddit communities, rather than a statistically representative sample of all LLM users.

Keyword Curation. We curate a keyword list dedicated to LLM privacy discussions by triangulating multiple resources (e.g., AI Privacy Taxonomy [42] [1] and the W3C Data Privacy Vocabulary v2.2 [31][2]). Furthermore, we incorporate keywords curated during the manual exploration of privacy policies described in Section 3.1.2. We remove overly generic terms (e.g., *use*) to avoid generating excessive irrelevant false positives during subsequent crawling processes. Additionally, we add LLM usage-context terms frequently appearing in privacy discussions (e.g., *chat history* and *dialogue log*). For multi-word keywords, we apply exact phrase matching. To improve recall, we expand each base keyword into common lexical and morphological variants (as in [26, 37]); for example, we include tense/derivation variants such as *collect/collected/collecting/collection* and *retain/retention*, and we include phrase variants such as *opt out/opt-out*. The final list contains 69 base terms and 251 total search variations, agreed upon by the research team.

Data Crawling. We collected Reddit data from the targeted subreddits in January 2026 through keyword-based searching. Following prior work [26, 44, 51], we query Reddit search results from the past year and retrieve up to the top 100 posts ranked by popularity (score). All duplicate posts across keywords are removed during the collection stage. During crawling, we skip records associated with moderators/bots or removed content (e.g., `AutoModerator`, `[deleted]`, `[removed]`) [26, 45]. We preserve necessary metadata fields, including `title`, `selftext` (post body), `body` (comment body), `score`, `parent_id`, `link_id`, `permalink`, and `num_comments` (as in [32]), which are essential for subsequent correlation analysis and review. We remove author identifiers to meet ethical requirements [28, 64]. Additional compliance considerations are detailed in Section *Ethical Considerations*. Overall, the collected corpus contains 28,306 threads and 512,550 comments, resulting in 540,856 records.

Data Preprocessing. We focus on threads that discuss privacy issues specific to the LLM product of each corresponding subreddit. To support scalable analysis, we design a two-phase preprocessing pipeline instead of relying on random or purely manual filtering. This approach adopts automated selection to increase data inclusiveness while simultaneously reducing human effort. First, we perform rule-based pre-filtering to remove low-signal threads, including posts with no human comments, records with non-positive score, and records with empty or deleted content. This step reduces the 28,306 threads to 19,774 candidate threads. Second, we apply an LLM-based relevance filter that classifies each candidate post as relevant or irrelevant based on its `title` and `selftext`. Unlike prior approaches requiring training a topic classifier with labeled data (e.g., DeBERTa-based filtering [45]), we leverage an LLM with few-shot learning capability and a strict, subreddit-specific relevance definition that distinguishes privacy from general security/safety issues. The filter removes obvious off-topic posts with minimal human involvement while preserving downstream manual review. The final filtered corpus contains 1,840 posts and 30,166 corresponding comments for subsequent analysis. These records undergo further coding, analysis, and manual review in later stages, as introducing LLM at this stage reduces human effort rather than replacing human judgment entirely.

Dataset Overview. To provide context for subsequent privacy concern extraction and gap analysis, we briefly summarize the distributional characteristics of the collected Reddit dataset (see Figure 3 and 4 in Appendix A). The final corpus comprises 1,840 posts and 30,166 comments, spanning 7 major LLM-related subreddits. The volume of posts varies significantly among communities; `r/ChatGPT` constitutes the majority (652 posts), followed by `r/GeminiAI` (309), `r/ClaudeAI` (248), and `r/OpenAI` (227). Conversely, `r/grok` (221), `r/Bard` (110), and `r/DeepSeek` (73) show lower activity levels. This disparity aligns with the market share and community maturity of the respective LLM providers; nevertheless, the dataset maintains sufficient diversity to represent perspectives across multiple major platforms. Keywords associated with user concerns cover diverse dimensions, including regulatory compliance (e.g., *Consent*, *GDPR [6]*), data security (e.g., *Authentication*, *Encryption*), and model behavior (e.g., *Training*, *Accuracy*). Keywords such as *memory* and *conversation data* appear frequently, reflecting the unique characteristics of LLM contexts.

3.1.2 Policy Document Collection

To maintain consistency with the Reddit data collection, we collect privacy policies for the 5 mainstream LLM products corresponding to the 7 subreddits, namely ChatGPT, Gemini, Claude, Grok, and DeepSeek. Our selection is smaller than that in [66] because we consider whether the corresponding subreddit has sufficient discussion activity. Moreover, this paper primarily contributes to gap analysis between user privacy concerns and privacy policies rather than comprehensive policy evolution analysis. For ChatGPT, since it is OpenAI's

[1] https://hankhplee.com/aiprivacytaxonomy/
[2] https://w3c.github.io/dpv/2.2/dpv/

core product and its privacy practices are directly governed by OpenAI's privacy policy, we treat OpenAI's privacy policy as the corresponding policy for ChatGPT. The same rationale applies to Claude, Grok, and DeepSeek in relation to their respective companies. Notably, following prior work [66], we exclude Google's general privacy policy. The portions of Gemini's privacy practices that follow Google's general policy are already cited and incorporated in Gemini-specific privacy documentation. Consequently, including the broader Google general policy would add little informational value while introducing substantial irrelevant content pertaining to services such as YouTube and Maps. Instead, we collect targeted supplementary documents, such as the Gemini Apps Privacy Hub, to ensure data relevance and specificity.

Document Identification and Acquisition. The collection of privacy policies follows the established methodology in prior work [26, 59, 66], employing manual browsing to locate target policy documents. In particular, we reference the approach used in [66] for collecting LLM provider privacy documents and use their curated and open-sourced privacy policy list as a baseline. We manually acquire the latest versions of privacy policies and supplementary privacy-related documents for these 5 LLM providers. After preliminary design and iterative discussion of the methodology, we obtained the most recent data before the final formal analysis and reporting, resulting in a collection date of January 21, 2026.

Data Processing and Overview. For each collected privacy policy document, we save both HTML and Markdown formats. The HTML format preserves the original page structure and formatting, facilitating traceability and reference. The Markdown format removes HTML markup, facilitating subsequent text analysis pipeline processing. Subsequently, researchers manually clean the Markdown files, removing boilerplate elements unrelated to the policy's substantive content, including navigation bars, headers/footers, and cookie consent banners. This preprocessing step ensures that the policy text used in subsequent gap analysis contains only actual privacy-related disclosures. Ultimately, our privacy policy dataset covers 5 LLM providers, comprising 33 privacy-related documents. Table 5 in Appendix A presents detailed information on the collected documents, including document names, word counts, and update dates.

3.2 Data Analysis

As discussed in Section 2, existing research on Reddit or privacy policy analysis typically relies on sampling-based, manual qualitative methods, which is generally acknowledged as a limitation in terms of scalability and reproducibility. Since recent studies [23, 46, 48, 55, 61, 71] successfully introduced LLMs into security and privacy research, we present an LLM-assisted semi-automated pipeline for large-scale privacy concern extraction and privacy policy gap analysis. Ethical considerations regarding the use of LLMs for Reddit data analysis are detailed in Section *Ethical Considerations*.

Specifically, we leverage LLMs to systematically extract privacy concerns from extensive Reddit discussions and audit them against detailed policy content. Furthermore, we evaluate the LLM-assisted pipeline through two complementary human validation procedures (Section 3.2.4): a pre-annotated ground truth that evaluates the performance of pipeline outputs, and an independent post-hoc double validation of LLM-generated gap records, following best practices in prior LLM-assisted research [11, 21, 57, 71].

3.2.1 Privacy Concern Extraction

By analyzing user discussions in the Reddit corpus, we gain deep insights into public privacy sentiment and focus areas regarding LLM services. This stage employs an LLM-assisted method to systematically extract privacy concerns and corresponding topics from each Reddit thread, establishing a foundation for subsequent gap identification. We process data at the thread level, where each thread contains one post and all its comments, enabling a more accurate understanding of the context surrounding each record.

Topic Taxonomy. To achieve a systematic analysis of user privacy concerns, we develop a taxonomy of privacy concern topics as guidance. The topic taxonomy evolves through an iterative process, with each iteration achieving consensus through research team meetings and discussions. First, the research team establishes an initial topic taxonomy based on privacy-related literature [31, 42] and preliminary data exploration. Subsequently, using this taxonomy as guidance, we employ the LLM to automatically categorize privacy concern topics for 100 randomly sampled records (discarding those without concerns). The research team convenes meetings after each iteration to review the LLM's classification results, discuss divergent cases, and reach consensus on adjustments to the topic taxonomy. These refinements mainly address overly narrow or ambiguous labels, overlaps between categories, and LLM-specific concerns insufficiently captured by general privacy taxonomies. After three rounds of iterative refinement, we establish a stable classification system of 20 topics, ensuring that the vast majority of concerns can be accurately categorized and maintaining good coverage (saturation) when new data is introduced, consistent with typical qualitative analysis coding processes [64]. The 20 privacy concern topics are organized into four main dimensions:

- **Group A: Data Lifecycle** (8 topics): Covers personal data collection (e.g., Input Content, Behavioral Metadata), usage (e.g., Service Provision, Model Training), retention and deletion, and sharing (e.g., Third-party Sharing, Plugin Access).
- **Group B: User Rights and Control** (4 topics): Covers consent mechanisms, granular control options, transparency disclosure, and policy change notification.
- **Group C: AI-specific Considerations** (4 topics): Covers

AI-specific privacy risks, including output privacy risks (data leakage, inference attacks), memory and personalization features, agent autonomous actions, and downstream integration scenarios (API, embedded AI).

- **Group D: Compliance and Protection** (4 topics): Covers jurisdiction and applicable laws (e.g., GDPR [6], CCPA [1]), vulnerable population protection (e.g., children, elderly), data security measures, and breach notification.

Complete topic definitions, descriptions, and examples are provided in our repository [3]. Moreover, during the LLM-assisted extraction process, we allow the LLM to cautiously propose new topics when the existing topic taxonomy provides insufficient coverage. These new proposals undergo human expert validation and research team deliberation before decisions are made regarding their inclusion or merger.

Extraction Process. For each thread, the LLM reviews the post and each comment (i.e., *record*) sequentially, identifying expressed privacy concerns. For each identified concern, the LLM outputs structured information, including but not limited to the assigned *topic* category, a clear *concern statement* articulating the user's privacy concern, a direct *supporting quote* from the original text, and brief *reasoning* explaining the topic selection to enhance interpretability. A thread may contain multiple concerns from different records, and a record may express multiple distinct concerns. The LLM is explicitly instructed to extract only content directly related to privacy and personal data, excluding general product quality complaints, technical issues, feature requests unrelated to privacy, or broad security issues, thereby ensuring the precision of extraction results.

3.2.2 Privacy Policy Gap Detection

Our method leverages LLM scalability to audit every individual user concern against the privacy policy, ensuring that no privacy policy gap is overlooked due to thematic abstraction, while maintaining traceability between each gap and its original user concern.

Gap Taxonomy. Following the same iterative procedure as the concern topic taxonomy, we construct the gap taxonomy from privacy policy literature, preliminary gap cases, and team deliberation. Refinements mainly address partial or ambiguous category names, overlaps between candidate gap types, and LLM-specific policy issues that do not fit traditional policy-analysis categories. We define 6 gap types organized into two categories (details are in our repository):

(1) *Privacy Policy Coverage Gaps* (4 types, objective policy content deficiencies that can be directly verified by reading policy text). These gaps require strict judgment criteria: the concern and topic must fall under legally mandated disclosure items, the policy must actually mention the topic, the language must be objectively vague, and the vagueness must prevent users from understanding their data rights.

- **Policy Detail Vague (G1)**: The policy mentions the relevant topic but uses vague language (e.g., "*may*", "*reasonable*", "*appropriate*") without providing specific details, failing to meet legal disclosure requirements.
- **AI Feature Unaddressed (G2)**: Privacy implications of AI-specific features (e.g., Memory/Personalization, Multimodal processing, Agent autonomous operations, Model Training data handling) are not covered in the policy.
- **Vulnerable Group Neglected (G3)**: Specialized privacy protection provisions for vulnerable populations (e.g., *children*, *elderly users*) are absent.
- **Jurisdiction Unclear (G4)**: Issues such as cross-border data transfers, applicable laws, and region-specific user rights (e.g., GDPR [6] rights for EU users) are not clearly explained.

(2) *Privacy User Perception Gaps* (2 types, user perception-level issues, premised on the policy already adequately covering the relevant topic).

- **Explicit Policy Distrust (G5)**: Although the policy adequately covers the relevant topic, users explicitly express distrust of policy statements (e.g., "*I don't believe they actually delete my data*").
- **Policy Awareness Deficit (G6)**: The policy actually covers the topic of user concern, but users are unaware of or unable to locate the relevant content, reflecting issues with policy discoverability or clarity.

This taxonomy is designed to be provider-agnostic rather than service-specific, thereby enhancing the comparability of analysis results. During LLM-assisted analysis, we similarly allow the LLM to cautiously propose new gap types when the existing taxonomy provides insufficient coverage. These proposals are reviewed by human experts and deliberated by the research team before decisions are made regarding their inclusion or merger.

Necessity Assessment and Legal Grounding. Our two RQs serve different purposes. For RQ1, every user-expressed privacy concern is a valid observation worth documenting. For RQ2, however, a concern should be counted as a policy gap only if it falls within a disclosure topic that a privacy policy is legally or normatively expected to address. Therefore, before gap classification, the LLM performs *Necessity Assessment* to determine whether the concern warrants privacy policy analysis. This step grounds the judgment in legal provisions and regulatory requirements, rather than subjective impressions, and filters ambiguous or mis-scoped concerns that would otherwise inflate the significance of detected gaps.

The legal grounding draws on major privacy and data-protection frameworks relevant to the global user base of the five providers, including GDPR, CCPA/CPRA, COPPA, HIPAA, the CLOUD Act, China's PIPL, and South Korea's PIPA. Note that state-of-the-art LLMs (e.g., GPT-5) have demonstrated substantial general legal knowledge and legal-reasoning capability in evaluations [24, 29, 54, 71]. These frameworks inform whether a concern implicates disclosure

[3] https://zenodo.org/records/20310585

duties such as categories of personal data, purposes of processing, retention criteria, data recipients, children's privacy, sensitive or health-related data, cross-border access, or jurisdiction-specific rights. For example, a concern such as "session tokens may not be rotated frequently enough, creating privacy exposure" is out of scope for RQ2 because token rotation is primarily an engineering and security-control issue, not a privacy policy disclosure obligation. Only when a concern involves a legally mandated or industry-standard disclosure topic does the LLM proceed to policy review and gap classification. The reliability of this step is evaluated through the validation procedures reported in Section 4.1.

Analysis Process. For each concern extracted in Section 3.2.1, the system provides the LLM with concern information (including concern statement, user assumption, supporting quote, and other structured information output from *Privacy Concern Extraction*) along with the corresponding provider's privacy policy text. The LLM first performs Necessity Assessment, determining whether the concern falls within the scope that privacy policies should cover based on legal and industry standards. If deemed necessary, the LLM searches the policy for relevant content and evaluates coverage status (Not Found / Vague / Adequate). Finally, based on the policy evidence found and legal reasoning, the LLM determines whether a gap exists and classifies it, outputting the gap detection result, coverage status, exact quotes from the policy as evidence, a legally-grounded justification, and preliminary recommendations for policy improvement.

Note that long or fragmented provider disclosures make policy review a more demanding evidence-localization task; recent LLM-based privacy policy analysis similarly improves review reliability by structuring policy inputs [71]. In our pipeline, we support reliable evidence tracing by auditing each concern separately against cleaned provider-specific Markdown policy text, requiring exact policy quotations and structured JSON reasoning, and validating sampled gap records through expert review.

3.2.3 Prompt Engineering Principles

When utilizing the LLM to complete the above analysis tasks, we follow several key prompt (i.e., natural language instructions for guiding the LLM to complete tasks) design principles to ensure output quality and consistency. First, we employ the Tree of Thoughts (ToT) [73] method to guide the LLM in classification decisions, requiring the LLM to generate multiple candidate options (e.g., topics, gaps), evaluate their degree of match sequentially, and then select the best result. This multi-path reasoning approach is more robust than linear thinking. Second, we decompose complex tasks into multiple sequential phases (e.g., gap detection includes four stages: *User Concern Understanding*, *Necessity Assessment*, *Policy Review*, and *Gap Classification*), ensuring the reasoning process is transparent and auditable. Moreover, we establish strict scope constraints in the prompts and explicitly list exclusion criteria to prevent the LLM from over-extracting non-privacy content. We also require all outputs to provide supporting evidence (e.g., exact quotes), enhancing result verifiability. All LLM outputs adopt a pre-defined JSON schema to ensure they are structured and machine-parseable. We additionally provide examples consistent with the output format to illustrate input cases, reasoning processes, and analysis results. Complete prompts are detailed in our repository.

Our pipeline is designed to support any state-of-the-art LLM, and GPT-5 is adopted as a representative instantiation in this study; other capable models could equally be substituted.

3.2.4 Expert Validation

To validate the effectiveness of the LLM-assisted analysis, we use two complementary human validation procedures, each involving two human experts, covering *Concern Extraction*, *Necessity Assessment*, and *Gap Detection*. Across the two procedures, all four experts have at least 8 years of cybersecurity research and practical experience. Before annotation or validation, each expert reviews all 33 collected privacy policy documents to ensure familiarity with provider-specific policy coverage.

Pre-annotated Ground Truth. To establish an independent baseline for pipeline evaluation, we randomly sample 110 complete threads from the preprocessed Reddit corpus. These threads contain 438 records across all five providers. Two experts annotate the sampled records using the established codebook (taxonomy). Following prior qualitative coding practice [11], the experts resolve disagreements through discussion and produce a consensus ground truth (intercoder agreement is therefore not reported separately). The resulting ground truth contains 226 records with privacy concerns (51.6%), yielding 295 concern instances (as one record may contain multiple concerns), of which 263 are deemed necessary for further policy analysis and 191 have identifiable gaps.

Post-hoc Independent Double Validation. The ground truth comparison measures how closely pipeline outputs match independent human annotations, including missed concerns and strict category alignment. However, it does not directly indicate whether each generated gap record is evidence-backed. Therefore, we conduct post-hoc validation to ensure that experts review the full structured reasoning chain (concern statement, supporting quote, retrieved policy excerpts, legal justification, and gap reasoning), rather than merely the final LLM outputs. Note that the two experts in *Pre-annotated Ground Truth* do not participate in post-hoc validation, thereby avoiding any potential bias resulting from duplicate reviews. We employ a *stratified random sampling* strategy to construct the validation sample set. Sample size determination is based on statistical principles: for the total of 3,137 gap instances identified (detailed in Section 4), achiev-

Table 1: Post-hoc independent double validation of 355 LLM-generated records.

Dimension	Agreement	Cohen's κ	Gwet's AC1	Final Accepted
Concern	88.45%	0.7285	0.7993	92.96%
Necessity	89.30%	0.7324	0.8217	95.49%
Gap	91.27%	0.7955	0.8477	92.11%

ing a $\pm 5\%$ margin of error at a 95% confidence level requires approximately 343 samples. Considering that LLM judgment accuracy for the same gap type may vary across different topics, we apply the following sampling rule for each *(Provider, Topic Type, Gap Type)* combination: if the number of gaps $N \leq 3$ for a given topic under the LLM, all are included in the validation set; if $N > 3$, we randomly sample 3[4]. For gap types and concern topics newly proposed by the LLM (identified with a "NEW_" prefix), since they are few in number and require research team deliberation to determine whether to incorporate them into the taxonomy or merge them with existing categories, we include all of them in the validation set. This sampling design yields 355 validation samples (11.3% of all gap instances).

Validation Interface and Adjudication. We develop a dedicated web-based interface to support efficient and structured expert review (see our repository). The interface presents each sampled gap with the full reasoning chain, including the original discussion context, concern statement, supporting quote, assigned topic, retrieved policy excerpts, legal justification, gap reasoning, and direct links to source materials where needed. Two experts independently review the same 355 LLM-generated records and mark each dimension as TRUE (agreeing with the LLM's reasoning) or FALSE (disagreeing with the LLM's reasoning or considering it debatable). Subsequently, the research team discussed the cases involving FALSE and reached a final decision.

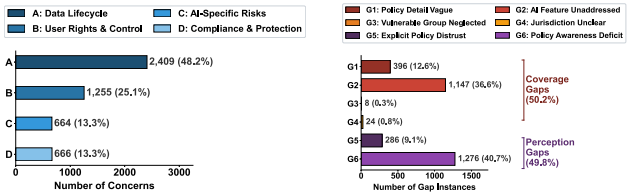
4 Result

4.1 Pipeline Validation

We report results from two complementary human validation studies, with detailed confusion matrices available in our repository.

Compared with the pre-annotated ground truth, the pipeline achieves an F1 score of 90.04% for concern extraction. For necessity assessment and gap-type detection, we report conditional metrics on aligned concern instances, because both labels are defined over extracted concerns. The pipeline achieves conditional F1 scores of 95.66% and 75.09%, respectively. The lower gap-type score reflects the more difficult

[4]We use 3 as the threshold because larger numbers would impose excessive burden on the expert, while smaller numbers would fail to meet the overall sample size requirement



(a) Distribution of extracted user privacy concerns (4,994).

(b) Distribution of identified privacy gaps (3,137).

Figure 1: Overview of large-scale privacy gap analysis results.

Table 2: Prevalence of privacy concerns and gaps.

Level	w/ Concern	w/ Gap	Gap | Concern
Thread (1,531)	1,054 (68.8%)	883 (57.7%)	83.8%
Record (8,527)	3,628 (42.5%)	2,381 (27.9%)	65.6%

Note: Privacy Concern and Policy Gap columns indicate the presence of at least one instance. Gap Probability represents the likelihood of a policy gap occurring when a concern is identified.

one-to-many setting, because a concern may map to multiple gap types and boundary cases may introduce new gap formulations.

For post-hoc validation, Table 1 reports the results for the sampled 355 LLM-generated records. Overall, it shows high inter-coder agreement and high final acceptance across all three dimensions. We report Gwet's AC1 alongside Cohen's κ because true/accepted instances dominate this setting; consequently, κ may understate agreement under class imbalance [13, 68]. For privacy concern analysis, our results align with prior LLM-assisted privacy work. For example, Ali et al. [11] report 96.8% accuracy and 91.2% F1 score using hierarchical LLM-based classification for privacy concern analysis. For gap detection, the 92.11% post-hoc final acceptance implies with 95% confidence that LLM reasoning accuracy falls within [89.47%, 94.75%], further demonstrating pipeline reliability.

Given the validated effectiveness of LLM-assisted analysis, the RQ1 and RQ2 analyses below report aggregate statistics over the full pipeline outputs. For newly identified concern topics and gap types not present in the initial taxonomy, the research team decided through discussion to merge each into the most relevant existing category, thereby avoiding an overly fragmented taxonomy while keeping the classification system compact. An overview of the distribution of extracted concerns and identified gaps is shown in Figure 1 and Table 2.

4.2 RQ1: Community-Expressed Concerns

We analyze 4,994 privacy concerns extracted from 1,531 threads across five LLM provider communities. These concerns are categorized into our 20-topic taxonomy spanning four groups: Data Lifecycle (Group A, 48.2%), User Rights

Table 3: Distribution of user privacy concerns across LLM platforms.

Privacy Topic Total (Concerns [Threads])	ChatGPT 3272 [879]	Claude 587 [248]	Gemini 282 [110]	Grok 710 [221]	DeepSeek 143 [73]
Group A: Data Lifecycle (2409/4994, 48.2%)					
A1.1 Input Content	256 (16.1%) [143 (16.3%)]	45 (17.2%) [30 (12.1%)]	18 (13.4%) [15 (13.6%)]	17 (4.8%) [17 (7.7%)]	24 (35.8%) [20 (27.4%)]
A1.2 Behavioral Metadata	182 (11.4%) [114 (13.0%)]	18 (6.9%) [13 (5.2%)]	11 (8.2%) [9 (8.2%)]	41 (11.6%) [27 (12.2%)]	10 (14.9%) [7 (9.6%)]
A2.1 Service Provision	53 (3.3%) [45 (5.1%)]	-	4 (3.0%) [4 (3.6%)]	13 (3.7%) [12 (5.4%)]	-
A2.2 Model Training	209 (13.1%) [144 (16.4%)]	73 (27.9%) [44 (17.7%)]	49 (36.6%) [27 (24.5%)]	31 (8.8%) [19 (8.6%)]	14 (20.9%) [9 (12.3%)]
A3.1 Retention Duration	313 (19.6%) [171 (19.5%)]	58 (22.1%) [38 (15.3%)]	39 (29.1%) [21 (19.1%)]	90 (25.6%) [46 (20.8%)]	8 (11.9%) [5 (6.8%)]
A3.2 Deletion Mechanism	274 (17.2%) [145 (16.5%)]	14 (5.3%) [10 (4.0%)]	7 (5.2%) [6 (5.5%)]	105 (29.8%) [39 (17.6%)]	1 (1.5%) [1 (1.4%)]
A4.1 Third Party Sharing	274 (17.2%) [122 (13.9%)]	35 (13.4%) [27 (10.9%)]	4 (3.0%) [3 (2.7%)]	48 (13.6%) [31 (14.0%)]	10 (14.9%) [8 (11.0%)]
A4.2 Plugin Extension Access	33 (2.1%) [19 (2.2%)]	19 (7.3%) [14 (5.6%)]	2 (1.5%) [2 (1.8%)]	7 (2.0%) [5 (2.3%)]	-
Group B: User Rights & Control (1255/4994, 25.1%)					
B1.1 Consent Mechanism	98 (12.6%) [76 (8.6%)]	75 (41.2%) [34 (13.7%)]	11 (11.6%) [9 (8.2%)]	41 (23.0%) [27 (12.2%)]	2 (8.3%) [2 (2.7%)]
B1.2 Granular Control	314 (40.5%) [210 (23.9%)]	51 (28.0%) [34 (13.7%)]	56 (58.9%) [31 (28.2%)]	65 (36.5%) [46 (20.8%)]	3 (12.5%) [2 (2.7%)]
B1.3 Transparency Disclosure	350 (45.1%) [224 (25.5%)]	54 (29.7%) [36 (14.5%)]	27 (28.4%) [18 (16.4%)]	72 (40.4%) [48 (21.7%)]	19 (79.2%) [11 (15.1%)]
B1.4 Policy Change Notification	14 (1.8%) [12 (1.4%)]	2 (1.1%) [2 (0.8%)]	1 (1.1%) [1 (0.9%)]	-	-
Group C: AI-Specific Risks (664/4994, 13.3%)					
C1.1 Output Risk	174 (38.4%) [118 (13.4%)]	24 (32.9%) [19 (7.7%)]	10 (33.3%) [7 (6.4%)]	55 (57.3%) [34 (15.4%)]	4 (33.3%) [3 (4.1%)]
C1.2 Memory Personalization	249 (55.0%) [145 (16.5%)]	14 (19.2%) [10 (4.0%)]	16 (53.3%) [11 (10.0%)]	33 (34.4%) [25 (11.3%)]	6 (50.0%) [5 (6.8%)]
C2.1 Agent Autonomous Actions	22 (4.9%) [17 (1.9%)]	26 (35.6%) [14 (5.6%)]	-	2 (2.1%) [2 (0.9%)]	1 (8.3%) [1 (1.4%)]
C2.2 Downstream Integration	8 (1.8%) [8 (0.9%)]	9 (12.3%) [9 (3.6%)]	4 (13.3%) [4 (3.6%)]	6 (6.2%) [6 (2.7%)]	1 (8.3%) [1 (1.4%)]
Group D: Compliance & Protection (666/4994, 13.3%)					
D1.1 Jurisdiction Law	149 (33.2%) [86 (9.8%)]	22 (31.4%) [15 (6.0%)]	8 (34.8%) [6 (5.5%)]	14 (16.7%) [12 (5.4%)]	19 (47.5%) [12 (16.4%)]
D1.2 Vulnerable Population	7 (1.6%) [5 (0.6%)]	-	-	1 (1.2%) [1 (0.5%)]	2 (5.0%) [2 (2.7%)]
D2.1 Data Security	268 (59.7%) [129 (14.7%)]	47 (67.1%) [33 (13.3%)]	15 (65.2%) [12 (10.9%)]	69 (82.1%) [29 (13.1%)]	16 (40.0%) [11 (15.1%)]
D2.2 Breach Notification	25 (5.6%) [19 (2.2%)]	1 (1.4%) [1 (0.4%)]	-	-	3 (7.5%) [2 (2.7%)]

Note: Highlighted cells indicate the top 2 concerns in each group. Format: Concerns (Group Share %) [Records (Thread Penetration %)].

& Control (Group B, 25.1%), AI-Specific Risks (Group C, 13.3%), and Compliance & Protection (Group D, 13.3%). Table 3 presents the distribution of concerns across topics and providers.

Data Lifecycle Practices Dominate User Concerns. Concerns related to data lifecycle practices (Group A) account for the largest share of user-expressed privacy concerns, representing 48.2% (2,409/4,994) of all extracted concerns. This finding aligns with prior survey research [11, 74], which identified data collection (43.7%) and training usage (22.5%) as users' primary concerns. The dominance of Group A is consistent across all five providers, with each provider's Group A concerns ranging from 44.6% to 49.6% of their total.

Among the eight topics in this group, Retention Duration (A3.1) emerges as particularly salient, ranking among the top two concerns within Group A for four out of five providers. Users express uncertainty about storage duration and whether deletion requests are honored. As one user articulated, *"I assumed my chat history would be protected under GDPR and that opting out would be sufficient. Now it seems all conversations may be retained indefinitely, potentially shared, and impossible to fully delete."*

Users also frequently raise concerns about Input Content (A1.1), Model Training (A2.2), and Deletion Mechanism (A3.2). For input content, users worry about inadvertent disclosure of sensitive information: *"From now on, avoid sharing sensitive information with Grok since xAI can access everything you provide."* Regarding model training, a common assumption persists: *"I figured they use whatever we provide to improve their models."* For deletion, users report frustration with interface changes: *"The option to clear conversation history seems to have disappeared from the settings."* These concerns collectively reflect users' desire for greater clarity and control over the complete data lifecycle.

Users Demand Transparency and Fine-Grained Control. User rights concerns (Group B) constitute 25.1% (1,255/4,994) of all concerns. Prior work [36] documented that LLM providers often scatter critical privacy information across multiple sub-policies and FAQ documents. Our analysis confirms this observation: Transparency Disclosure (B1.3) emerges as the most frequent concern in this group, accounting for 45.1% (350/776) of ChatGPT's Group B concerns and reaching 79.2% (19/24) for DeepSeek. Users express frustration when privacy-relevant information is not surfaced prominently. One user noted, *"The App Store's privacy section for xAI only mentions collecting identifiers and diagnostics, but xAI's full privacy policy buried on their website reveals that all chat conversations are stored to improve services. This lack of upfront clarity feels deceptive."* Users also express concerns about dynamic privacy setting changes without adequate notice: *"They can toggle features like the memory system on or off at will. Nothing is communicated transparently."*

Granular Control (B1.2) represents another significant concern category, comprising 40.5% (314/776) of ChatGPT's Group B concerns. Users seek fine-grained mechanisms to manage their data, particularly regarding training opt-outs and conversation-level controls: *"Does disabling 'Improve the model for everyone' in settings achieve the same privacy*

protection as using temporary chat mode?" Such questions reveal confusion about the relationship between different privacy controls and their actual effects.

AI-Specific Risks Raise Emerging Concerns. AI-Specific Risks (Group C) account for 13.3% (664/4,994) of concerns. As Tao et al. [66] document, LLM privacy policies often fail to address novel AI capabilities, instead relying on traditional software privacy language. Our findings reveal that users express concerns about AI-specific privacy challenges that traditional frameworks may inadequately address.

Output Risk (C1.1) emerges as users' major AI-specific concern, accounting for 38.4% (174/453) of ChatGPT's Group C concerns and 57.3% (55/96) for Grok. Users worry about LLMs generating content that reveals personal information: *"Grok pulled his facial features from photos he had posted on X and generated images that looked strikingly similar to him."* Such concerns about generative capabilities represent a novel privacy challenge distinct from traditional data collection issues. Memory Personalization (C1.2) constitutes the largest concern in this group for ChatGPT, representing 55.0% (249/453) of its Group C concerns. Users express uncertainty about cross-session information retention: *"I started a fresh conversation and asked about my preferences, and it already knew, despite nothing being visible in the persistent memory settings I could access."* This disconnect between user expectations and observed system behavior underscores the opacity of AI memory mechanisms.

Data Security and Jurisdictional Ambiguity Cause Anxiety. Compliance & Protection concerns (Group D) represent 13.3% (666/4,994) of the total, with Data Security (D2.1) and Jurisdiction Law (D1.1) as the most prominent topics. D2.1 dominates this group across most providers, accounting for 59.7% (268/449) of ChatGPT's Group D concerns and reaching 82.1% (69/84) for Grok. For data security, users express concerns about authentication tokens and session management: *"Third-party integrations could potentially exploit your OAuth tokens or API keys to extract data from your account. Go to OpenAI's website, navigate to Settings, then Security, and log out of all devices to reset compromised tokens."* Such technical awareness suggests a subset of users possess sophisticated understanding of potential attack vectors.

Jurisdictional concerns are particularly pronounced for providers with international operations. This pattern is especially evident for DeepSeek, where 47.5% (19/40) of its Group D concerns focus on jurisdiction: *"DeepSeek stores data on servers in China, raising concerns about potential access by the Chinese government."* Another user advised: *"By running DeepSeek locally, you eliminate the risk of sending your data to Chinese servers."* These responses illustrate how geopolitical considerations increasingly shape users' privacy decisions.

Provider-Specific Patterns Emerge. While concern distributions show broad consistency across providers, notable provider-specific patterns emerge. ChatGPT yields the highest volume with 3,272 concerns from 879 threads, reflecting its larger user base and more active communities (3.7 concerns per thread versus 2.0–3.2 for others). DeepSeek triggers heightened jurisdictional concerns, with Jurisdiction Law (D1.1) accounting for 47.5% of its Group D concerns as users express apprehension about cross-border data transfers to China. Claude users focus on Consent Mechanism (B1.1) at 41.2% of Group B, questioning consent practices in autonomous AI interactions. Grok users raise prominent Deletion Mechanism (A3.2) concerns at 29.8% of Group A, particularly regarding image generation outputs. Gemini users emphasize Granular Control (B1.2) at 58.9% of Group B, the highest among all providers, likely reflecting concerns about integration with Google's broader ecosystem.

4.3 RQ2: Policy-Concern Gaps

Building on the 4,994 user concerns identified in RQ1, we assess whether providers' privacy policies adequately address these concerns. Our LLM-assisted gap detection pipeline identifies 3,137 gap instances across six categories. These gaps fall into two broad classes: *policy coverage gaps* (G1–G4, 50.2%) where policies fail to address user concerns, and *user perception gaps* (G5–G6, 49.8%) where policies provide coverage but users remain unaware or distrustful. Table 4 and Figure 2 present the distribution.

Policy Coverage and User Perception Gaps Are Equally Prevalent. Our analysis reveals a near-equal split between policy-side and user-side gaps. Policy coverage gaps (G1–G4) account for 50.2% (1,575/3,137) of all identified gaps, while user perception gaps (G5–G6) constitute 49.8% (1,562/3,137). This balanced distribution suggests that improving LLM privacy requires addressing both inadequate policy disclosures and user comprehension barriers.

Among policy coverage gaps, AI Feature Unaddressed (G2) dominates at 36.6% (1,147/3,137), indicating that novel AI capabilities frequently lack policy coverage. Detail Vague (G1) accounts for 12.6% (396/3,137), where policies use vague language that fails to specify concrete practices. Vulnerable Group Neglected (G3) and Jurisdiction Unclear (G4) are less frequent at 0.3% (8/3,137) and 0.8% (24/3,137) respectively, but carry significant regulatory implications. For user perception gaps, Policy Awareness Deficit (G6) is the most prevalent single gap category at 40.7% (1,276/3,137), exceeding all coverage gap types. Explicit Policy Distrust (G5) represents 9.1% (286/3,137). The high prevalence of G6 aligns with prior findings that LLM privacy policies are longer, more complex, and more ambiguous than traditional software policies [66].

Vague Disclosures and Unaddressed AI Features Dominate Coverage Gaps. G1 (Detail Vague) gaps indicate that policies acknowledge a privacy topic but provide insufficient detail. From Table 4, Retention Duration (A3.1) exhibits high G1 counts, reflecting user frustration with ambiguous retention disclosures. Under GDPR Article 13(2)(a), controllers

Table 4: Cross-category mapping of privacy topics to gap classifications across LLM providers.

Privacy Topic	Gap Categories (*Provider order:* ChatGPT Claude Gemini Grok DeepSeek)																													
	G1 Policy Detail Vague					**G2** AI Feature Unaddressed					**G3** Vulnerable Group Neglected					**G4** Jurisdiction Unclear					**G5** Explicit Policy Distrust					**G6** Policy Awareness Deficit				
A1.1 Input Content	44	5	1	1	0	32	6	1	1	0	0	0	0	0	0	0	0	0	0	0	15	0	0	1	0	51	12	6	4	9
A1.2 Behavioral Metadata	8	0	0	0	1	16	0	0	3	1	1	0	0	0	0	0	0	0	0	0	4	1	0	1	0	73	8	3	13	1
A2.1 Service Provision	3	0	0	0	0	29	0	1	6	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	0	9	0	1	4	0
A2.2 Model Training	4	2	1	0	0	34	5	14	4	1	0	0	0	0	0	0	0	0	0	0	24	7	0	2	1	89	26	12	16	6
A3.1 Retention Duration	152	8	3	27	0	40	2	0	2	0	0	0	0	0	0	0	0	0	0	0	6	2	0	8	0	54	22	10	27	2
A3.2 Deletion Mechanism	12	0	0	4	0	119	1	0	9	1	0	0	0	0	0	0	0	0	0	0	34	2	1	18	0	49	5	3	50	0
A4.1 Third Party Sharing	7	0	2	3	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	28	1	0	2	0	115	13	0	20	3
A4.2 Plugin Extension Access	0	0	2	0	0	19	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0
B1.1 Consent Mechanism	5	4	1	6	1	24	8	1	4	0	0	0	1	0	0	1	0	0	0	1	3	7	1	1	0	21	18	1	14	0
B1.2 Granular Control	1	0	1	0	0	93	7	5	14	1	1	0	1	0	0	0	0	0	0	0	12	2	1	2	0	82	23	24	17	2
B1.3 Transparency Disclosure	20	3	1	4	4	98	5	11	14	5	0	0	0	0	0	0	0	1	0	1	25	7	0	9	2	108	28	9	35	6
B1.4 Policy Change Notification	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
C1.1 Output Risk	5	1	0	0	0	110	5	10	28	4	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	14	3	0	12	0
C1.2 Memory Personalization	0	0	0	0	0	240	14	6	28	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	4	3	0
C2.1 Agent Autonomous Actions	0	0	0	0	0	11	17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	3	0	0	1
C2.2 Downstream Integration	0	1	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	3	0
D1.1 Jurisdiction Law	21	1	0	0	15	5	1	1	1	1	0	0	0	0	0	10	1	6	0	3	9	3	0	1	0	29	4	0	5	0
D1.2 Vulnerable Population	0	0	0	0	0	1	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D2.1 Data Security	3	0	0	3	0	4	1	0	2	0	0	0	0	0	0	0	0	0	0	0	24	4	0	6	1	33	16	2	15	2
D2.2 Breach Notification	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

must disclose the period for which personal data will be stored or, if that is not possible, the criteria used to determine that period. Additionally, Article 13(1)(c) requires disclosure of processing purposes and legal basis. Yet policies often state only general criteria without specifying durations. One user expressed concern: *"According to Anthropic's documentation, if a prompt triggers safety classifiers, inputs and outputs may be retained for up to two years. Does this apply even to paid accounts?"*

G2 (AI Feature Unaddressed) gaps are the largest category, representing 36.6% of all gaps. These gaps concentrate in Group C (AI-Specific Risks), particularly Memory Personalization (C1.2) and Output Risk (C1.1). Users frequently encounter situations where policy opt-outs do not apply to specific AI features: *"I disabled training data usage, but ChatGPT still asks me to choose between two responses for A/B testing. What happens to those choices?"* The concentration of G2 gaps reflects that LLM privacy policies often borrow language from traditional software disclosures [66], leaving novel AI capabilities without adequate coverage.

Regulatory-Sensitive Gaps Affect Specific Populations. Although less frequent, Vulnerable Group Neglected (G3) and Jurisdiction Unclear (G4) gaps carry significant regulatory implications. G3 gaps (8/3,137, 0.3%) arise when policies fail to address protections for children and other vulnerable groups. Parents express concerns: *"The feature was not blocked for my preschooler. I received an email notification and immediately checked—it was already enabled on her tablet account."* Under COPPA [4] (15 U.S.C. §6501 & §6502), verifiable parental consent is required before collecting personal information from children under 13.

G4 gaps (24/3,137, 0.8%) occur when policies fail to clarify applicable legal frameworks, particularly for cross-border data or health-related services. In G4, the vantage point and jurisdictional scope depend on the potential situations and specific jurisdictions presented in each context. One user questioned: *"You overestimate how much HIPAA protects you here. I don't think HIPAA would even apply to ChatGPT Health."* When services invite users to connect medical records, reasonable expectations about HIPAA protections arise. Yet policies often fail to clarify whether the provider operates as a covered entity under 45 C.F.R. §160.103.

Awareness Deficits Outpace Explicit Distrust. User perception gaps constitute nearly half of all identified gaps (49.8%, 1,562/3,137), indicating that policy coverage alone is insufficient to address user concerns. Policy Awareness Deficit (G6) is the largest single gap category at 40.7% (1,276/3,137). G6 gaps occur when policies address concerns but users remain unaware of the coverage. One user explained: *"I am not comfortable using DeepSeek's browser version as it explicitly states that it trains the model on my inputs. I don't think I can opt out."* User awareness deficits stem from two sources: users often skim policies superficially [74], and LLM policies are longer and more difficult to read than traditional software policies [66].

Explicit Policy Distrust (G5) represents 9.1% (286/3,137)

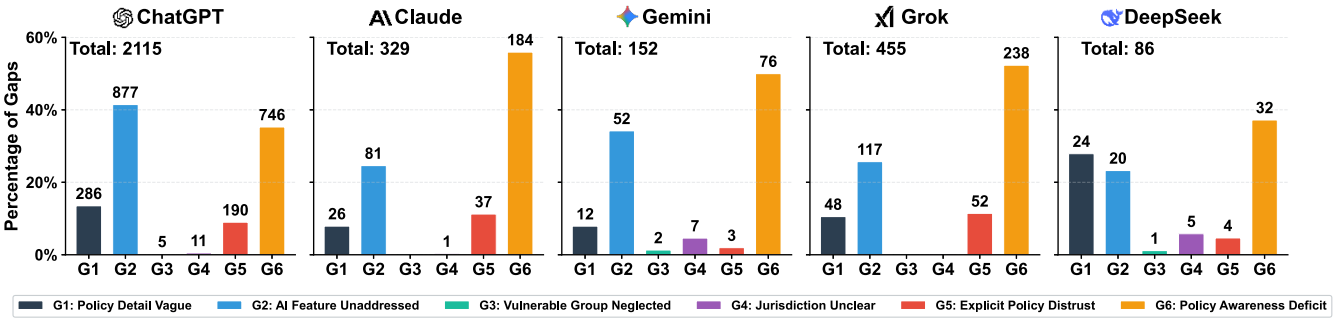


Figure 2: Privacy gap distribution across different providers.

of gaps, where users explicitly distrust provider claims regardless of policy content. One user stated: *"The problem is that ChatGPT was specifically programmed to lie about this aspect of its data collection, which is concerning."* Such statements reflect fundamental credibility issues that policy amendments alone cannot resolve, necessitating public communication or technical verification to restore user trust.

Gap Patterns Vary Across Providers. Gap distributions reveal distinct patterns across providers, as shown in Figure 2. ChatGPT accounts for 67.4% (2,115/3,137) of all gaps, reflecting its larger user base. G2 (AI Feature Unaddressed) is particularly prominent for ChatGPT at 41.5% (877/2,115), likely reflecting its extensive feature set including memory, plugins, and multimodal capabilities. DeepSeek shows elevated G1 (Detail Vague) proportion at 27.9% (24/86), indicating insufficient detail on cross-border data handling. Claude exhibits notable Distrust (G5) concerns at 11.2% (37/329) of its total gaps. Grok and Gemini show dominant Awareness Deficit (G6) proportions at 52.3% (238/455) and 50.0% (76/152) respectively, suggesting users of these platforms may be less familiar with available policy disclosures.

Potential Discrepancies Between Policy Claims and System Behavior. Beyond policy coverage gaps, our analysis surfaced user reports of potential inconsistencies between stated policies and actual system behavior. While we cannot independently verify all claims, these reports warrant attention. One user documented a concerning observation about ChatGPT's training opt-out: *"While checking the browser Inspector tool, I clicked on a JSON response that loads when opening Settings > Data Controls. I've had 'Improve the model for everyone' set to OFF for over a year. But in the network inspector, I noticed: data_usage_for_training: 'permitted'. Even though the toggle is OFF."* The user provided steps to reproduce this observation and noted that toggling the setting did not change the underlying value. Note that confirming the above impactful inconsistencies requires system-level access and provider cooperation, which falls beyond our scope. If confirmed, such discrepancies would represent a more severe issue than disclosure gaps, highlighting the need for verifiable privacy guarantees rather than policy promises alone. On the other hand, such cases illustrate why policy review alone is insufficient for rapidly evolving LLM services, where new settings and features may change faster than disclosures or user-facing explanations.

5 Discussion

Our analysis of 3,137 privacy gap instances across five major LLM providers reveals systematic disconnects between user concerns and policy disclosures. This section interprets our findings and discusses implications for providers, users, and policymakers.

5.1 Key Insights and Implications

Our gap taxonomy reveals that LLM privacy challenges differ fundamentally from traditional software privacy issues. The concentration of gaps in AI-specific categories—particularly awareness deficits (40.7%) and unaddressed features (36.6%)—suggests that existing privacy frameworks struggle to address conversational AI's unique characteristics. Coverage gaps (G1–G4) and perception gaps (G5–G6) each account for approximately half of all identified gaps (50.2% vs. 49.8%), indicating that policy inadequacy and user understanding deficits are equally significant challenges. Although awareness deficits parallel findings in other domains, our study reveals novel LLM-specific taxonomies, cross-provider patterns unattainable by prior studies, and actionable stakeholder insights specific to LLM privacy gaps.

Awareness Gaps Dominate Despite Policy Coverage. The largest gap category, Policy Awareness Deficit (G6, 40.7%), indicates that policy coverage alone is insufficient. Users frequently express concerns about topics that policies technically address but fail to surface effectively. For instance, training opt-out mechanisms exist across all providers, yet users regularly ask whether their data trains models and how to prevent it. This pattern has regulatory implications: GDPR [6] Article 12 mandates "transparent, intelligible, and easily accessible"

disclosures, while CCPA [1] requires information to be "reasonably accessible." Our findings suggest that providers may satisfy the letter of transparency requirements while undermining their spirit through scattered, lengthy documentation. Prior work confirms this fragmentation [36]; LLM providers distribute critical privacy information across multiple subpolicies, FAQs, and help articles.

AI-Specific Features Outpace Policy Disclosures. AI Feature Unaddressed (G2, 36.6%) represents the second-largest gap category, concentrated in memory, personalization, and output-related concerns. All five providers exhibit similar patterns: policies describe traditional data collection but rarely address how AI-specific capabilities—including cross-session memory, context accumulation, and model memorization risks—affect user privacy [38, 62]. ChatGPT shows the highest G2 proportion (41.5%), reflecting its extensive feature set including persistent memory, custom GPTs, and third-party plugins. Users express confusion about whether deleting a conversation removes data from memory, whether earlier messages can be "forgotten" mid-session, and how cross-chat personalization works. This lag between feature deployment and policy adaptation aligns with Tao et al.'s [66] observation that LLM policies often borrow language from traditional software disclosures without substantive modification. Moreover, newly introduced capabilities create coverage gaps before policies catch up, while older but poorly explained features continue to generate awareness deficits and distrust.

Vagueness as Strategic Ambiguity. Detail vague gaps (G1, 12.6%) are concentrated primarily around retention duration, with additional cases appearing in input content handling, jurisdictional disclosures, transparency, and deletion mechanisms. Policies often rely on retention "criteria" under GDPR [6] Article 13(2)(a) rather than fixed timelines, which may formally follow the structure of the provision but still leaves users without operational clarity. Phrases like "as long as we need" or "for our legitimate business purposes" appear across providers, offering legal compliance without user understanding. This vagueness may reflect genuine operational uncertainty—LLM providers cannot always predict how long training data remains influential within model weights. However, providers could distinguish between data categories (conversation logs, memory items, account metadata) and provide concrete timelines where possible. DeepSeek shows the highest G1 proportion (27.9%), potentially reflecting its newer market entry and less developed policy infrastructure.

Provider-Specific Patterns Reveal Market Dynamics. Gap distributions vary meaningfully across providers, reflecting different market positions and feature sets. ChatGPT accounts for 67.4% (2,115/3,137) of all gaps with G2 at 41.5%, driven by its extensive feature ecosystem including memory, plugins, and custom GPTs. Claude, Grok, and Gemini exhibit elevated awareness deficit proportions, with G6 accounting for 55.9%, 52.3%, and 50.0% of their gaps, respectively, suggesting policy information exists but remains poorly surfaced. DeepSeek shows the highest G1 proportion at 27.9%, reflecting user uncertainty about data practices from a China-headquartered provider [67]. Explicit distrust (G5) is relatively elevated for both Grok and Claude, accounting for 11.4% and 11.2% of their provider-specific gaps, respectively. This suggests that explicit distrust is not confined to a single provider and may reflect broader user skepticism toward whether stated privacy commitments are implemented in practice. Taken together, these provider-level profiles indicate that mitigation should be aligned with each platform's dominant gap mechanism, prioritizing feature-specific disclosures for high-G2 ecosystems, improved policy discoverability for high-G6 providers, clearer data-practice specifications for high-G1 cases, and verifiable assurances where G5 is pronounced.

Explicit Distrust Signals Deeper Concerns. Explicit Policy Distrust (G5, 9.1%) represents users who express skepticism that providers follow their stated policies regardless of content. This category, while smaller, indicates fundamental trust erosion that improved disclosures alone cannot address. User quotes reveal the nature of this distrust: concerns about discrepancies between stated deletion capabilities and observed system behavior, suspicion that training opt-outs are not honored, and belief that legal compliance claims are performative rather than substantive. One user noted observing a discrepancy between the UI toggle state and underlying API values for training opt-out settings. These perceptions, whether verified or not, highlight the need for verifiable privacy guarantees rather than policy promises.

5.2 Recommendations for Stakeholders

The following recommendations primarily reflect the comprehensive analysis of gap distributions, policy evidence, user concern examples, preliminary recommendations from LLM outputs, and related work.

Closing Disclosure Gaps: For LLM Providers. *Surface AI-Specific Feature Disclosures.* The prevalence of G2 gaps (36.6%) indicates systematic under-disclosure of AI-specific capabilities. Providers should add dedicated policy sections addressing: (1) memory and personalization features with retention periods and deletion controls, (2) context window handling clarifying that mid-session deletion requests cannot remove earlier messages from active processing, and (3) output-side risks explaining safeguards against model memorization and cross-user data leakage. *Specify Concrete Retention Periods.* Vague retention language generates substantial user confusion. Providers should specify retention timelines for distinct data categories including conversation logs, memory items, and training datasets. *Improve Policy Accessibility.* Awareness deficits (40.7%) suggest that information exists but remains undiscoverable. Providers should surface critical disclosures through contextual notices, privacy dashboards, and searchable FAQ systems.

Navigating the Current Landscape: For Users. *Adopt a*

Privacy-First Interaction Mindset. Users should treat LLM interactions as semi-public communications rather than private conversations. Even with training opt-outs enabled, conversation data may persist in logs, backups, and safety review systems for extended periods. *Leverage Temporary and Ephemeral Modes.* Where available, temporary chat modes offer meaningfully different retention characteristics. Users handling sensitive topics should default to these modes rather than relying on after-the-fact deletion requests. *Monitor Provider Communications.* Privacy practices change frequently in response to new features and regulatory developments. Users should periodically revisit privacy settings as default configurations evolve.

Strengthening Governance: For Regulators. *Mandate AI-Specific Disclosure Requirements.* Existing frameworks such as GDPR [6] and CCPA/CPRA [1, 2] establish general transparency obligations but do not address AI-unique scenarios including memory personalization, context window handling, and model memorization risks. Regulators should develop supplementary guidance clarifying how existing requirements apply to conversational AI. *Require Standardized Privacy Summaries.* Following mobile ecosystem precedents where iOS and Android introduced standardized privacy labels, regulators could mandate comparable disclosures for AI services covering training data usage, retention periods, and memory features. *Establish Compliance Verification Mechanisms.* User distrust (G5, 9.1%) reflects skepticism about whether providers follow stated policies. Because LLM products and feature sets evolve rapidly, compliance verification should be periodic and feature-triggered rather than one-time. Regulators should explore third-party audit requirements or certification schemes that compare policy claims with privacy settings, API/network behavior, and retention/deletion evidence after major feature releases.

Enabling Technical Accountability: For Tool Developers. *Design Input Sanitization Tools.* User concerns about sensitive data exposure suggest need for client-side utilities that detect and redact personally identifiable information before submission, operating as browser extensions or API wrappers. *Build Policy Comparison Interfaces.* The awareness deficit gap (40.7%) suggests users struggle to locate relevant policy information. Developers could create tools that aggregate and present LLM privacy policies in comparable formats, highlighting key differences. *Develop Verification Mechanisms.* User distrust stems partly from inability to verify provider claims. Technical approaches such as traffic analysis tools or API behavior monitors could provide users with independent evidence about actual data practices.

5.3 Limitations and Future Work

We acknowledge several limitations suggesting directions for future research.

Scope and Generalizability. Our data collection spans Reddit communities, whose users tend to be more technologically engaged than the general population. Consistent with related Reddit-based security and privacy studies [28, 37, 44, 45, 64], keyword search and score-ranked retrieval may overrepresent active, high-salience, or privacy-motivated discussions. Our seven-subreddit, five-provider design mitigates single-community or single-provider selection bias, but the results should be interpreted as structural gap signals rather than population-prevalence estimates. Privacy expectations also vary across cultural and regulatory contexts. Future work should examine concerns across broader populations, including enterprise and API users who operate under different privacy regimes with contractual protections.

Temporal and Methodological Constraints. LLM privacy policies evolve rapidly; our analysis provides only a snapshot at the current point in time. Additionally, our analysis focuses on policy documents and does not include privacy setting interfaces, where users might seek information beyond policies and where providers may expose more detailed privacy controls. Our concern extraction also relies on self-reported experiences rather than verified system behavior, though our gap taxonomy explicitly distinguishes coverage gaps from perception gaps to address this limitation.

Policy-Behavior Verification. Our analysis identifies user reports of potential discrepancies between stated policies and observed behavior—including deleted memories that reappear and training opt-outs that seem ineffective. Without technical access to service providers' systems and their cooperation in disclosure, such reports remain difficult to verify. Systematic policy-behavior verification through network analysis, API auditing, or controlled experiments represents a critical research direction that could transform user perceptions into validated findings.

6 Conclusion

In this paper, we pave a new avenue to understand whether LLM privacy policies adequately address what users genuinely care about. To achieve this goal, we first systematically extract privacy concerns from Reddit communities of five major LLM providers, developing a 20-topic taxonomy organized into four thematic groups. Moreover, we assess these concerns against providers' latest privacy policies, constructing a 6-type gap classification that distinguishes policy coverage deficiencies from user perception barriers. Our analysis of 4,994 user-expressed concerns and 3,137 identified gaps reveals that AI-specific feature gaps (36.6%) and user awareness deficits (40.7%) together comprise the vast majority (77.3%) of all gaps, while policy-side and user-side problems contribute nearly equally (50.2% vs. 49.8%). Notably, we also observe user-reported evidence of discrepancies between stated privacy policies and actual system behavior, highlighting the urgent need for verifiable privacy guarantees and technical accountability in LLM services. Overall, our

work demonstrates distinct novelty in several aspects: the **Problem**—bridging the gap between policy evaluation and user concerns, which prior research has examined independently but not in conjunction; the **Approach**—presenting the first large-scale, user-centered policy audit pipeline that systematically maps community-expressed concerns to policy disclosures, distinguishing it from traditional policy-only or survey-based methods; and the **Results**—providing both quantitative and qualitative insights into critical disclosure failures and perception barriers for this important yet underexplored problem. We believe our work lays the foundation for bridging the disconnect between LLM privacy policies and user expectations, offering actionable insights for relevant stakeholders to advance toward user-centered privacy design and verifiable privacy guarantees in AI systems.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62472434) and the Key Program of NSFC Hunan (2026JJ30028).

Ethical Considerations

Our institution's Institutional Review Board (IRB) reviewed this project and determined it to be exempt from full review, as it involves analysis of publicly available data without direct interaction with human subjects.

This study analyzes publicly posted Reddit content and privacy policy documents from LLM providers. To minimize potential privacy risks, we implemented several measures. (1) *Data de-identification.* We remove all author identifiers (usernames) before analysis and do not attempt to re-identify or contact any users. (2) *Presentation safeguards.* When presenting illustrative examples in this paper, we paraphrase content rather than quoting verbatim and adjust wording to prevent direct search, following established practices in social media research [25, 26, 32]. (3) *Data retention.* The aggregated dataset is not released publicly, and our analysis focuses on the substance of expressed concerns rather than the individuals who expressed them.

Regarding LLM-assisted analysis, we use commercial LLM APIs for automated content analysis, consistent with recent work in social media research [11, 39, 57]. We adhere to provider data processing terms: OpenAI's API policy states that data sent to the API is not used to train or improve models, and zero data retention endpoints do not store submitted content.

Overall, the stakeholder-specific ethical considerations are as follows. For Reddit users, the primary risk is unwanted identifiability or contextual exposure, which we mitigate through de-identification, paraphrasing, and withholding the raw corpus. For LLM providers, the risk is over-attributing user-reported concerns to verified system behavior; therefore, we analyze publicly available policy disclosures, report aggregate patterns, and avoid claims of provider misconduct without independent verification. For downstream users, regulators, and researchers, the potential benefit is improved transparency and accountability, while the main risk is overgeneralization beyond the sampled communities; accordingly, we frame findings as policy-concern gap signals and document limitations in Section 5.3.

Open Science

We release the relevant artifacts associated with this paper, including the LLM-related privacy keywords for data collection, the collected policy datasets with preprocessing results, the analysis scripts, the developed taxonomies, the prompts used as LLM instructions, and the web-based interface for validation. These artifacts are accessible at https://zenodo.org/records/20310585 and https://github.com/TheFatInsect/LLM-Privacy-Gap. Due to the aforementioned ethical considerations, we do not disclose the collected Reddit corpus or its corresponding processing results.

References

[1] California consumer privacy act. https://oag.ca.gov/privacy/ccpa.

[2] The california privacy rights act of 2020. https://thecpra.org/.

[3] Chatgpt. https://chatgpt.com.

[4] Children's online privacy protection rule. https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

[5] Claude. https://claude.ai.

[6] General data protection regulation. https://gdpr-info.eu/.

[7] Google gemini. https://gemini.google.com.

[8] Samsung bans staff's ai use after spotting chatgpt data leak. *The Straits Times*, 2023.

[9] March 20 chatgpt outage: Here's what happened. https://openai.com/index/march-20-chatgpt-outage/, 2024.

[10] State of llm security report 2025 | cobalt. https://resource.cobalt.io/state-of-llm-security, 2025.

[11] Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. Understanding users' security and privacy concerns and attitudes towards conversational ai platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 298–316, 2025.

[12] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 187–195, 2021.

[13] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I. Bertenthal, and Apu Kapadia. Influencing photo sharing decisions on social media: A case of paradoxical findings. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1350–1366, 2020.

[14] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. Policylint: Investigating internal privacy policy contradictions on google play. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 585–602, 2019.

[15] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. How knowledge workers use and want to use llms in an enterprise context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[16] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2824–2843, 2021.

[17] Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-Jun Li, and Yaxing Yao. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 277–297, 2025.

[18] Avishek Choudhury and Hamid Shamszare. Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis. *Journal of Medical Internet Research*, 25(1):e47184, 2023.

[19] Sadia Sultana Chowa, Riasad Alvi, Subhey Sadi Rahman, Md Abdur Rahman, Mohaimenul Azam Khan Raiaan, Md Rafiqul Islam, Mukhtar Hussain, and Sami Azam. From language to action: a review of large language models as autonomous agents and tool users. *Artificial Intelligence Review*, 59(2):71, 2026.

[20] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. Poligraph: Automated privacy policy analysis using knowledge graphs. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1037–1054, 2023.

[21] Fernando M. Delgado-Chaves, Matthew J. Jennings, Antonio Atalaia, Justus Wolff, Rita Horvath, Zeinab M. Mamdouh, Jan Baumbach, and Linda Baumbach. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences*, 122(2):e2411962122, 2025.

[22] Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143, 2025.

[23] Chongzhou Fang, Ning Miao, Shaurya Srivastav, Jialin Liu, Ruoyu Zhang, Ruijie Fang, Asmita, Ryan Tsang, Najmeh Nazari, Han Wang, and Houman Homayoun. Large language models for code analysis: Do llms really do their job? In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 829–846, 2024.

[24] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, 2024.

[25] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah Gilbert, and Naiyan Jones. Remember the human: A systematic review of ethical considerations in reddit research. *Proc. ACM Hum.-Comput. Interact.*, 8(GROUP):5:1–5:33, 2024.

[26] Lan Gao, Oscar Chen, Rachel Lee, Nick Feamster, Chenhao Tan, and Marshini Chetty. "i cannot write this because it violates our content policy": Understanding content moderation policies and user experiences in generative ai products. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3727–3746, 2025.

[27] Martyna Gliniecka. The ethics of publicly available data research: A situated ethics framework for reddit. *Social Media + Society*, 9(3):20563051231192021, 2023.

[28] Maggie Yongqi Guan, Yaman Yu, Tanusree Sharma, Molly Zhuangtong Huang, Kaihua Qin, Yang Wang, and Kanye Ye Wang. Security perceptions of users in stablecoins: Advantages and risks within the cryptocurrency ecosystem. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 42–42, 2024.

[29] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279, 2023.

[30] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, 2018.

[31] Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Paul Ryan, Delaram Golpayegani, and Julian Flake. Data privacy vocabulary (dpv) – version 2.0. In *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings, Part III*, pages 171–193, 2024.

[32] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 'i've talked to chatgpt about my issues last night.': Examining mental health conversations with large language models through reddit analysis. *Proc. ACM Hum.-Comput. Interact.*, 9(7):CSCW356:1–CSCW356:25, 2025.

[33] Emiram Kablo and Patricia Arias-Cabarcos. Privacy in the age of neurotechnology: Investigating public attitudes towards brain data collection and use. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–238, 2023.

[34] Dilara Kekulluoglu and Yasemin Acar. "we are a startup to the core": A qualitative interview study on the security and privacy development practices in turkish software startups. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2015–2031, 2022.

[35] Hanna Kim, Minkyoo Song, Seung Ho Na, Seungwon Shin, and Kimin Lee. When llms go online: The emerging threat of web-enabled llms. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 1729–1748, 2025.

[36] Jennifer King, Kevin Klyman, Emily Capstick, Tiffany Saade, and Victoria Hsieh. 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(2):1465–1477, 2025.

[37] Jan H. Klemmer, Stefan Albert Horstmann, Nikhil Patnaik, Cordelia Ludden, Cordell Burton, Carson Powers, Fabio Massacci, Akond Rahman, Daniel Votipka, Heather Richter Lipford, Awais Rashid, Alena Naiakshina, and Sascha Fahl. Using ai assistants in software development: A qualitative study on security practices and concerns. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2726–2740, 2024.

[38] Alexandra Klymenko, Stephen Meisenbacher, Patrick Gage Kelley, Sai Teja Peddinti, Kurt Thomas, and Florian Matthes. "we are not future-ready": Understanding ai privacy risks and existing mitigation strategies from the perspective of ai developers in europe. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, pages 113–132, 2025.

[39] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[40] Deepak Kumar, Jeff Hancock, Kurt Thomas, and Zakir Durumeric. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, pages 2797–2807, 2023.

[41] Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. Exploring user security and privacy attitudes and concerns toward the use of general-purpose llm chatbots for mental health. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6007–6024, 2025.

[42] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[43] Leon Leibmann, Galen Weld, Amy X. Zhang, and Tim Althoff. Reddit rules and rulers: Quantifying the link between rules and perceptions of governance across thousands of communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1098–1121, 2025.

[44] Jiliang Li, Nora Sinong Lu, Isaak Hanimann, Janice Jianing Si, Dazhao Cheng, Xiaobo Zhou, and Kanye Ye Wang. Investigating the impact of online community involvement on safety practices and perceived risks among people who use drugs. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6025–6044, 2025.

[45] Jingjie Li, Kaiwen Sun, Brittany Skye Huff, Anna Marie Bierley, Younghyun Kim, Florian Schaub, and Kassem Fawaz. "it's up to the consumer to be smart": Understanding the security and privacy attitudes of smart home users on reddit. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2850–2866, 2022.

[46] Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Hoon Wei Lim, and Bryan Hooi. Knowphish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 793–810, 2024.

[47] Fengyu Liu, Yuan Zhang, Jiaqi Luo, Jiarun Dai, Tian Chen, Letian Yuan, Zhengmin Yu, Youkun Shi, Ke Li, Chengyuan Zhou, Hao Chen, and Min Yang. Make agent defeat agent: Automatic detection of taint-style vulnerabilities in llm-based agents. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3767–3786, 2025.

[48] Peiyu Liu, Junming Liu, Lirong Fu, Kangjie Lu, Yifan Xia, Xuhong Zhang, Wenzhi Chen, Haiqin Weng, Shouling Ji, and Wenhai Wang. Exploring chatgpt's capabilities on vulnerability management. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 811–828, 2024.

[49] Yupei Liu, Yuqi Jia, Jinyuan Jia, and Neil Zhenqiang Gong. Evaluating llm-based personal information extraction and countermeasures. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 1669–1688, 2025.

[50] Zhihuang Liu, Ling Hu, Tongqing Zhou, Yonghao Tang, and Zhiping Cai. Prevalence overshadows concerns? understanding chinese users' privacy awareness and expectations towards llm-based healthcare consultation. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2716–2734, 2025.

[51] Rijul Magu, Nivedhitha Mathan Kumar, Yihe Liu, Xander Koo, Diyi Yang, and Amy Bruckman. Understanding online discussion across difference: Insights from gun discourse on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–28, 2024.

[52] Jaron Mink, Hadjer Benkraouda, Limin Yang, Arridhana Ciptadi, Ali Ahmadzadeh, Daniel Votipka, and Gang Wang. Everybody's got ml, tell me what else you have: Practitioners' perception of ml-based security tools and explanations. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2068–2085, 2023.

[53] Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dana Calacci, and Andrés Monroy-Hernández. Quallm: An llm-based framework to extract quantitative insights from online forums. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1355–1369, 2025.

[54] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report, 2024.

[55] Rupam Patir, Qiqing Huang, Keyan Guo, Wanda Guo, Guofei Gu, Haipeng Cai, and Hongxin Hu. Towards llm-assisted vulnerability detection and repair for open-source 5g ue implementations. In *Proceedings 2025 Workshop on Security and Privacy of Next-Generation Networks*, 2025.

[56] Elvira Pollina and Alvise Armellini. Italy fines openai over chatgpt privacy rules breach. *Reuters*, 2024.

[57] Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Thematic-lm: A llm-based multi-agent system for large-scale thematic analysis. In *Proceedings of the ACM on Web Conference 2025*, pages 649–658, 2025.

[58] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343, 2019.

[59] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. "community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2024.

[60] Juliane Schmüser, Philip Klostermeyer, Kay Friedrich, and Sascha Fahl. "i'm pretty expert and i still screw it up": Qualitative insights into experiences and challenges of designing and implementing cryptographic library apis. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2322–2340, 2025.

[61] Ronal Singh, Shahroz Tariq, Fatemeh Jalalvand, Mohan Baruwal Chhetri, Surya Nepal, Cecile Paris, and Martin Lochner. Llms in the soc: An empirical study of human-ai collaboration in security operations centres. *arXiv preprint arXiv:2508.18947*, 2025.

[62] Ilan Strauss, Isobel Moure, Tim O'Reilly, and Sruly Rosenblat. Real-world gaps in ai governance research. *arXiv preprint arXiv:2505.00174*, 2025.

[63] Tao Sun, Jian Xu, Yuanpeng Li, Zhao Yan, Ge Zhang, Lintao Xie, Lu Geng, Zheng Wang, Yueyan Chen, Qin Lin, Wenbo Duan, Kaixin Sui, and Yuanshuo Zhu. Bitsai-cr: Automated code review via llm in practice. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 274–285, 2025.

[64] Madiha Tabassum, Alana Mackey, Ashley Schuett, and Ada Lerner. Investigating moderation challenges to combating hate and harassment: The case of mod-admin power dynamics and feature misuse on reddit. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 37–54, 2024.

[65] Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7211–7219, 2025.

[66] Zhen Tao, Shidong Pan, Zhenchang Xing, Emily Black, Talia Gillis, and Chunyang Chen. A longitudinal measurement of privacy policy evolution for large language models. *arXiv preprint arXiv:2511.21758*, 2025.

[67] Aditi Uberoi. Deepseek cyber attack: Timeline, impact, and lessons learned. https://www.cm-alliance.com/cybersecurity-blog/deepseek-cyber-attack-timeline-impact-and-lessons-learned, 2025.

[68] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(1):61, 2013.

[69] Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing llm reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28489–28503, 2025.

[70] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[71] Qinge Xie, Karthik Ramakrishnan, and Frank Li. Evaluating privacy policies under modern privacy laws at scale: An llm-based automated approach. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 5797–5816, 2025.

[72] Jia Xu, Weilin Du, Xiao Liu, and Xuejun Li. Llm4workflow: An llm-based automated workflow model generation tool. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2394–2398, 2024.

[73] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[74] Shuning Zhang, Haobin Xing, Xin Yi, and Hewu Li. Natural language but omitted? on the ineffectiveness of large language models' privacy policy from end-users' perspective. *arXiv preprint arXiv:2406.18100*, 2024.

[75] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

A Dataset Overview

Table 5: Overview of collected privacy policy documents across major LLM providers. Coverage includes primary policies and supplementary disclosures, ensuring comprehensive representation of provider transparency practices as of January 2026. In the "Latest" and "Type" columns, filled circles (•) denote documents identified as the current/latest versions and primary policy materials, respectively; open circles (○) indicate materials whose latest-version status could not be verified or documents treated as supplementary disclosures, such as FAQs, addenda, terms, and related support pages.

Words	Document Name	Date	Latest	Type
Grok				
4,491	xAI privacy policy	Jul. 10, 2025	•	•
860	xAI Enterprise FAQs	Feb. 25, 2025	•	•
2,384	xAI Trust Statement	–	○	○
1,668	xAI's Europe Privacy Policy Addendum	Apr. 24, 2025	•	•
DeepSeek				
4,212	DeepSeek privacy policy	Dec. 22, 2025	•	•
Gemini				
1,696	Gemini Apps Privacy Hub	Jan. 21, 2026	•	•
2,331	Google API Services User Data Policy	Feb. 15, 2024	•	○
ChatGPT				
4,064	App Developer Terms	Dec. 17, 2025	•	○
3,127	OpenAI privacy policy	Jun. 27, 2025	•	•
3,388	OpenAI EU privacy policy	Nov. 04, 2024	•	•
2,290	Enterprise privacy at OpenAI	Jun. 04, 2025	•	○
3,008	OpenAI Data Processing Addendum	Dec. 01, 2025	•	○
702	How your data is used to improve model performance	Jan. 09, 2026	•	○
Claude				
2,650	Non-User Privacy Policy	Aug. 28, 2025	•	○
5,346	Anthropic privacy policy	Jan. 12, 2026	•	•
491	Deleting Claude accounts	Updated over a week ago	•	○
614	Sharing and Unsharing Chats	Updated over a week ago	•	○
557	How long do you store my data	Updated over a week ago	•	○
198	How can I export my Claude data?	Updated over a week ago	•	○
465	Is my data used for model training	Updated over a week ago	•	○
604	Can you delete data that I sent via API	Updated over a week ago	•	○
294	Does Anthropic Act as a Data Controller	Updated over a week ago	•	○
635	Does Anthropic crawl data from the web?	Updated over a week ago	•	○
353	Who owns and manages the data of my team?	Updated over a week ago	•	○
237	Can you delete data that I sent via Claude	Updated over a week ago	•	○
299	How can I delete or rename a conversation?	Updated over a week ago	•	○
1,001	How Do You Use Personal Data in Model Training?	Updated over a week ago	•	○
79	What is your approach to GDPR or related issues?	Updated over a week ago	•	○
338	What personal data will be processed by Computer use?	Updated over a week ago	•	○
366	How does Anthropic protect the personal data of Claude users?	Updated over a week ago	•	○
340	How does Clio analyze usage patterns while protecting user data?	Updated over a week ago	•	○
201	What Personal data is collected when using dictation on the Claude Mobile Apps?	Updated over a week ago	•	○
181	What is Anthropic's policy for handling governmental requests for user information?	Updated over a week ago	•	○

This appendix provides a multidimensional validation of the datasets underpinning our analysis, focusing on both policy documents and user discussion corpora.

Table 5 summarizes the composition and recency of collected privacy policy materials across major LLM providers, ensuring transparency of document coverage. Significant differences exist in the number and comprehensiveness of documents across providers. Anthropic provides the most extensive privacy-related documentation (20 documents), with its help center covering detailed explanations on specific topics such as data retention, model training, data deletion, and GDPR [6] compliance. In contrast, DeepSeek provides only one comprehensive privacy policy, which nevertheless encompasses a "Supplemental Clause - Jurisdiction-Specific" for the European Economic Area (EEA), Switzerland, and the UK (with 4,212 words). This difference reflects both the strategic variations among providers in privacy disclosure granularity and transparency, and affects the completeness of policy coverage in subsequent *Gap Analysis*. This study considers the imbalance in the number of privacy policies among different LLM providers to be a valuable finding. It reveals the heterogeneity in privacy information disclosure practices among current LLM providers, consistent with phenomena observed in prior work [66]. Moreover, since our gap analysis focuses on the mismatch between user concerns and policy disclosures, even if a provider offers only one policy document, we can still determine whether that policy adequately addresses users' specific concerns.

Figure 3 visualizes the prevalence of privacy-related keywords across different Reddit communities, revealing thematic density and cross-platform diversity in user concerns. Figure 4 complements this view by characterizing engagement patterns within the Reddit communities, including score distributions, comment activity, and community participation. Engagement metrics further reveal a heavy-tailed pattern: most posts receive modest attention, but a minority attract disproportionately high scores and comment counts. The average post score is 53.7, the mean number of comments per post is 23.5, and the average comment score is 6.9. These statistics indicate that, while broad participation is present, substantive discussion is concentrated in a subset of highly visible threads. However, we intentionally retain these lower-engagement threads (because privacy concerns are often specific, contextual, or technical [12, 32]). Together, these materials establish the representativeness, structural diversity, and analytical reliability of the datasets used in subsequent privacy concern extraction and policy gap analysis.

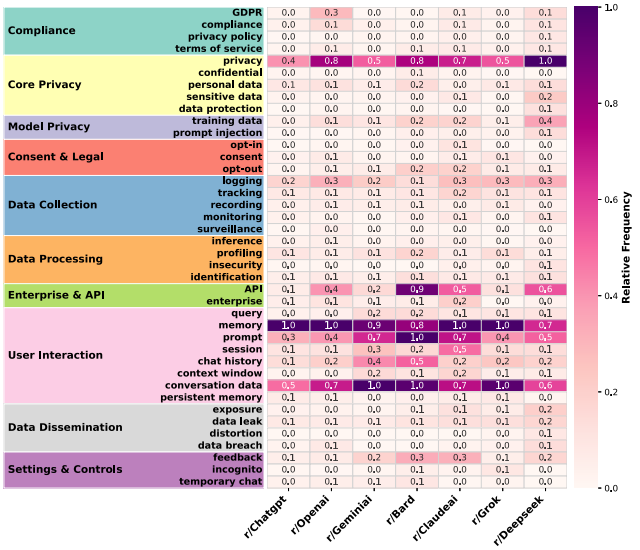
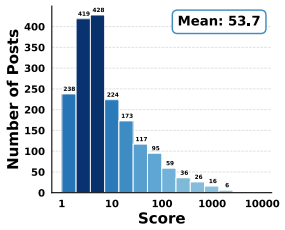
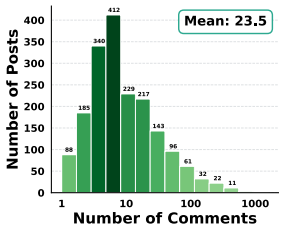


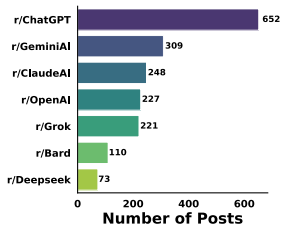
Figure 3: Heatmap of privacy keyword prevalence across LLM discussion communities. Relative keyword frequencies illustrate how privacy-related themes are distributed across Reddit communities associated with different LLM providers. The visualization highlights topical concentration, cross-platform consistency, and community-specific emphasis in user privacy discourse.



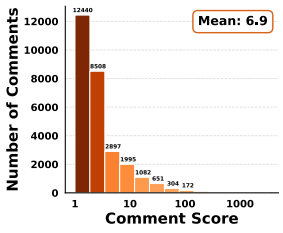
(a) Distribution of Post Scores



(b) Distribution of Comments per Post



(c) AI Community Posts



(d) Distribution of Comment Scores

Figure 4: Engagement characteristics of privacy-related discussions in the Reddit dataset.