

Chatbot Confessions: Large-Scale Analysis of Private Data Disclosure in Shared AI Chatbot Conversations

Majid Mollaefar
Center for Cybersecurity, Fondazione
Bruno Kessler (FBK)
Trento, Italy
mmollaefar@fbk.eu

Dimitri Van Landuyt
LIRIS, FEB, KU Leuven
Leuven, Belgium
dimitri.vanlanduyt@kuleuven.be

Gertjan Franken
DistriNet, KU Leuven
Leuven, Belgium
gertjan.franken@kuleuven.be

Nico Ebert
ZHAW, School of Management and
Law
Winterthur, Switzerland
ebet@zhaw.ch

Silvio Ranise
Center for Cybersecurity, Fondazione
Bruno Kessler (FBK)
Trento, Italy
ranise@fbk.eu

Abstract

The proliferation of AI conversation platforms has introduced unprecedented privacy risks through user-shared conversations. This paper presents a comprehensive analysis of privacy vulnerabilities in shared conversations across three major LLM platforms: ChatGPT, Microsoft Copilot, and Google Gemini. We collected and analyzed 100 342 conversations using an automated LLM-based privacy detection pipeline enhanced with a defined risk scoring system and the LINDDUN threat modeling framework. Our analysis identifies 8 131 conversations (8%) to incur privacy risks deriving from the disclosure of private and sensitive data including user identifiers (49%) and user location data (40%), yet in some cases also financial (4%), health (3%) and authentication data such as access tokens (3%). Through systematic analysis of conversation length and temporal disclosure patterns, we demonstrate that extended conversations exhibit higher privacy risk rates compared to brief interactions. Notably, 60% of private data disclosures in longer conversations occur in the final quartile of these conversations, which may indicate that users progressively lose privacy awareness as interactions deepen. Our findings have immediate implications for platform designers and policymakers, highlighting the need for proactive interventions including real-time privacy warnings, pre-share scanning, and clearer education about the permanence and discoverability of shared conversation links.

Keywords

AI Privacy, Conversational AI, LLMs, Personal Identifiable Information, Privacy Risk Analysis

1 Introduction

The rapid adoption of Large Language Models (LLMs) has fundamentally transformed human-AI interaction, with chatbot platforms such as OpenAI ChatGPT, Microsoft Copilot, and Google Gemini currently facilitating billions of conversational exchanges.

These chatbots are designed and optimized to deliver a user-friendly interface so that users can interact with technology in everyday situations by utilizing natural language. However, because of the nature of the data exchanged, and the possibility of abuse, these interactions raise privacy concerns for users [15]. Furthermore, research [9] indicates that the ability to construct natural language, to adopt human-like interaction patterns, and to generate high-quality responses also lead to reduced privacy vigilance compared to traditional web services [25]. This phenomenon is particularly concerning given the “intangible vulnerability” observed in users of mental health chatbots, where individuals underestimate the long-term privacy implications of their disclosures despite recognizing the sensitivity of the information [25]. The conversational nature of LLM interactions may encourage progressive disclosure, where users gradually reveal more sensitive information as conversations extend. This creates privacy risks when conversations intended as private exchanges are made publicly accessible through sharing mechanisms—a fundamental *contextual mismatch* between the intimate nature of conversational disclosure and permanent web publication.

This work focuses on privacy risks arising from private data disclosures in publicly shared chat conversations. These privacy risks stem from two contributing factors: (i) *private disclosure*—users sharing private information during conversational interactions with AI chatbots, and (ii) *public access*—platforms enabling public accessibility of these conversations through link sharing mechanisms. Both actions jointly create the privacy risk conditions under study.

The primary objective of this work is thus to increase our understanding of privacy risks in shared chatbot conversations, and to raise general awareness on these issues. We raise the following three research questions:

- RQ1.** What is the degree of private data disclosure in chatbot conversations?
- RQ2.** To which extent can these disclosures be considered high-risk and problematic from a privacy harm perspective?
- RQ3.** How do private disclosure patterns vary across conversation length and temporal progression?

To address these questions, we conduct an empirical analysis of private data disclosure in shared conversations from three major chatbot platforms: OpenAI ChatGPT, Microsoft Copilot, and Google

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.



Proceedings on Privacy Enhancing Technologies YYYY(X), 1–19
© YYYY Copyright held by the owner/author(s).
<https://doi.org/XXXXXXXX.XXXXXXX>

Gemini. These platforms provide a sharing feature that generates a unique link, allowing users to share their conversations with others. In August 2025, we collected 102 627 such conversations that had been indexed by mainstream search engines and preserved in public web archives [39]. We retain the user-provided content from these conversations by removing all AI responses, system messages, and other metadata, and apply a two-stage hybrid evaluation methodology:

The first stage (**Deterministic Pre-annotation**) involves applying specialized regular expressions (regex patterns) and validators to detect and annotate structured PII elements such as emails, phone numbers, social security numbers (SSNs), etc. This deterministic step enhances reproducibility, establishes stable baselines, and significantly reduces the risk of false positives common to purely generative detectors. This first stage also involves replacing each recognized PII element with cryptographically-hashed pseudonymous placeholders, the first in a series of tailored countermeasures and safeguards to prevent the potential amplification or exacerbation of privacy harms to affected data subjects.

In the second stage (**LLM-Driven Contextual Analysis**), we task an LLM to perform context-aware privacy analysis over the pre-annotated conversation text by applying a comprehensive privacy risk assessment framework. The LLM is provided with instructive prompts to: (i) identify private data disclosures beyond these fixed regex patterns (e.g., contextual health mentions, professional affiliations, and implicit location references), (ii) map these to the LINDDUN privacy threat type taxonomy [37], (iii) reason over combinations of quasi-identifiers to assess cumulative identifiability risk, and finally, (iv) assign multi-dimensional risk scores for the dimensions of sensitivity, identifiability, and exposure.

Our study results indicate that approximately 8% of the evaluated conversations incur privacy risks stemming from private data disclosures. Furthermore, we show that these disclosures are disproportionately more prevalent in the last quartile of the conversations, with notable concentration in later conversation stages.

This study provides the following three contributions:

- (1) **Scale:** Our study includes samples from three major chatbot platforms. This is to our knowledge the most complete evaluation of real-world AI chatbot conversations. Our findings raise concerns and demonstrate real-world implications for both users and providers of conversational chatbots.
- (2) **Automated processing pipeline:** We present and adopt a novel methodology for the structured, automated and reliable processing of textual chat conversations at scale. This processing pipeline is hybrid and two-staged, and combines deterministic detection precision with the contextual reasoning capabilities of LLMs. Furthermore, this evaluation pipeline has been validated for reproducibility and correctness against a ground truth created with expert annotators, achieving 77.5% accuracy with substantial inter-rater agreement (Fleiss' $\kappa = 0.799$), as detailed in Section 7.2.2.
- (3) **Temporal and length-based disclosure patterns:** We provide statistical evidence that privacy risk correlates significantly with conversation length, where large conversations ($> 15k$ characters) exhibit 2.5 times higher privacy risk rates compared to short conversations. Via temporal analysis, we show that

60% of private disclosures occur in the final quartile and reveal non-linear escalation patterns.

This paper is structured as follows. Section 2 provides background motivation for the work. Then, Section 3 introduces the hybrid privacy framework used for evaluating privacy risks. Section 4 then presents the overall methodology and adopted empirical approach. Section 5 presents the results, and Section 6 further discusses and validates our findings. Section 8 discusses related work. Finally, Section 9 concludes the paper.

2 Study Motivation: Privacy Risks of Chatbot Conversation Sharing

Zhao et al. [48] have published WildChat, a dataset consisting of 1 million ChatGPT conversations obtained with explicit user consent. Prior to dataset publication, Named Entity Recognition (NER) techniques were used to redact the dataset and reduce privacy harms. However, as demonstrated by Miresghallah et al. [31], these conversation traces still involve many additional in-context private data disclosures. The WildChat dataset, while large and extensive, was collected by providing users with unconstrained, anonymous usage. This biased the dataset towards a certain user demographic and disproportionately includes toxic chat conversation behavior.

A feature recently introduced in various chatbot platforms allows users to generate persistent public URLs for their chat conversations, providing public access to transcripts of entire conversations. This feature is intended for sharing, to support collaboration and reproducibility, and to facilitate overall knowledge dissemination. However, its initial introduction was characterized by limited user awareness and minimal privacy guidance. Security researchers discovered that conversations were also indexed by mainstream search engines, and preserved in public web archives [39]. Following user backlash and public controversy [8], platforms initiated de-indexing efforts to remove shared links from search results [13].

Motivation. The emerging practice of chatbot conversation sharing transforms ‘private’ user-chatbot conversations into durable web artifacts subject to large-scale collection and analysis. Such new features break an implicit expectation of user-chatbot confidentiality, and the uncontrolled sharing of conversation traces is particularly concerning given that private data disclosure in the course of a conversation does not only encompass direct identifiers (e.g., names, email addresses, phone numbers) but also quasi-identifiers (e.g., locations, dates, organizational affiliations) that become identifying when combined [40]. Despite initial efforts such as WildChat, a real-world evaluation of the incidence and nature of private data disclosure in shared chatbot conversations, and the sophisticated privacy risks and harms thereof, is currently lacking.

3 Privacy Risk Assessment Framework

We establish a comprehensive privacy risk assessment framework to systematically assess privacy risks in shared online conversations. This framework is designed to accommodate LLM-based evaluation, and is a key element of the overall study methodology. It integrates three core components: (i) a *privacy taxonomy* enumerating 12 major categories of privacy-relevant attributes with 44 granular codes; (ii) mappings to the LINDDUN *privacy threats* to connect the observed disclosures to concrete privacy harms; and (iii) a *risk*

scoring rubric that quantifies privacy risks along three dimensions (sensitivity, identifiability, and exposure) with explicit escalation rules. We present these components in the following subsections.

3.1 Privacy Taxonomy

The privacy taxonomy enumerates privacy-relevant attributes (PII codes) relevant to the context of chat conversations. Table 1 provides the overview, and shows the hierarchical nature of the taxonomy, consisting of 44 detailed PII codes, grouped into 12 higher-level categories. These detailed codes serve as the foundation for our automated privacy detection pipeline (Section 4.2.1).

Table 1: Taxonomy categories and associated PII codes.

Category	PII codes
CONTACT	EMAIL, PHONE, ADDRESS, SOCIAL_HANDLE
PROFILE	NAME, NAME_PUBLIC, DEMOGRAPHIC, FAMILY, MINOR, PROFESSIONAL
ID	GOVT, ACCOUNT_ID
BIO	FACE, FINGERPRINT, VOICE
AUTH	PASSWORD, API_KEY, TOKEN, SECRET
NET	IP_PUBLIC, IP_PRIVATE, MAC, HOSTNAME, URL_PRIVATE
FIN	CARD, BANK_ACCOUNT, TAX, CRYPTO_WALLET, TRANSACTION
HEALTH	CONDITION, MEDICATION, RECORD, MENTAL_HEALTH
LOC	ADDRESS, GPS, TRAVEL_HISTORY
CODE	PATH_LOCAL, REPO_PATH, BUILD_ID, CONTAINER_ID
LEGAL	CASE_NUMBER
SAFETY	VICTIM_TRAUMA, TRAFFICKING, MISSING_CHILD

Each category is expressly aligned to definitions from privacy engineering guidelines, standards and regulation (notably, NIST SP 800-122 [30], GDPR [3], ISO/IEC 29100 [2] and 27701 [4], NIST Privacy Framework [5]), and incorporates sectoral regimes where applicable (PCI DSS [6], HIPAA [1]). More details of this overall alignment and underpinning of the taxonomy can be found in Appendix A.

Below, we define each category precisely, indicate typical signals in chat conversations, and highlight intrinsic sensitivity per category¹.

CONTACT. Contact attributes encompass direct identifiers such as names, addresses, email, etc. that can directly identify or enable communication with specific individuals.

PROFILE. Profile attributes (names, demographic traits, job titles, employer, education history, family status, age bands, gender) refer to direct and quasi-identifiers that, in combination, can be used to re-identify individuals [3, 30]. In chats, PROFILE might appear in CV edits, personal introductions, or HR workflows; the re-identification risk compounds when multiple quasi-identifiers co-occur (e.g., ZIP+age+employer), consistent with classical results on linkage attacks [40].

ID. Identification numbers include government-issued identifiers such as Social Security Numbers (SSN), passport numbers, driver’s license numbers, and other official registry identifiers that uniquely identify individuals within administrative or governmental systems. These are considered highly sensitive PII.

¹Note that our taxonomy does not exhaustively enumerate all privacy-sensitive attributes; notably, GDPR Article 9 special categories such as religion, race, political opinion, and sexual orientation are not explicitly modeled as distinct categories. We excluded these categories because such personal disclosures are rarely direct, yet occur implicitly through context. Determining if conversations reveal, for example, political views is a subjective assessment that is more difficult to reliably and objectively assess.

BIO. Biometric identifiers can be used for uniquely and definitively identifying a person (e.g., fingerprints, faceprints, etc.). In chat contexts, references may appear as textual mentions (e.g., ‘my fingerprint scan failed’) or as metadata of attached images or enrollment flows; even textual descriptions can implicate biometric processing if linked to identification. Our risk scoring framework distinguishes abstract mentions (“my fingerprint scan failed”—low risk) from actual biometric data sharing (fingerprint images, voice recordings—high risk) through the Exposure dimension (see Section 3.3). Because biometrics lead to strong identification outcomes, we default to high sensitivity.

AUTH. While not always “personal data” in isolation, authentication and authorization artifacts—passwords, API keys, OAuth tokens/JWTs, cloud access keys, signing keys, and tokenized/signed URLs—are treated as high-risk information.

NET. These include network-layer and service identifiers (public/private IP addresses, hostnames, internal domains, private endpoints, and signed or tokenized URLs), which are often treated as personal data when they can identify or single out a user/device session. In practice, chats may reveal NET-category artifacts through logs, diagnostic output, or share-links with embedded access tokens. Risks include detectability (exposing internal surfaces to scanning), linkability (cross-artifact correlation via stable endpoints), and even direct disclosure when tokens grant access.

FIN. Financial identifiers include payment card data, bank account (IBAN, SWIFT) and routing numbers, tax identifiers, and crypto wallet addresses. These are explicitly covered across core references, and may lead to fraud and identity theft.

HEALTH. Health-related content (conditions, treatments, diagnoses, medications, mental-health references) constitutes sensitive personal data. Shared chats may include symptom descriptions, lab values, or provider interactions. Such data remains highly sensitive and potentially stigmatizing. Detection relies on medical lexicons and context cues; we prioritize confidentiality and minimization in processing.

LOC. Location data ranges from full postal addresses to GPS coordinates, time-stamped travel itineraries, and routine patterns (home-work commutes). In chats, LOC emerges via deliveries, travel planning, or debugging geo-services; even approximate regions can aid linkage across artifacts.

CODE. Code and infrastructure breadcrumbs (repository URLs, commit SHAs, internal paths, container image tags, CI/CD build IDs, environment variable dumps) can expose system internals and, indirectly, secrets or personal data flows. While primarily a security and confidentiality concern, CODE disclosures may lead to privacy harms when they enable identity impersonation (e.g., exposed developer credentials) or reveal personal data processing flows. Their disclosure is material to privacy-by-design and confidentiality (ISO/IEC 27701 governance). In chats, these may arise from troubleshooting and code reviews.

LEGAL/SAFETY. Legal identifiers (case numbers, docket references) and safety-sensitive contexts (e.g., victimization, trafficking, missing minors) warrant heightened care due to potential harm, stigma, and legal process risks. In chats, these can appear in appeals,

Table 2: Mapping taxonomy categories to LINDDUN threats.

Category	L	I	N	D	D	U	N
CONTACT	✓	✓	✗	✗	✓	✗	✗
PROFILE	✓	✓	✗	✗	✓	✗	✗
ID	✓	✓	✗	✗	✓	✓	✗
BIO	✓	✓	✗	✗	✓	✗	✗
AUTH	✓	✗	✗	✓	✓	✗	✓
FIN	✓	✓	✗	✗	✓	✗	✓
HEALTH	✗	✗	✗	✗	✓	✓	✓
LOC	✓	✗	✗	✓	✓	✗	✗
CODE	✓	✓	✗	✓	✓	✗	✓
NET	✓	✗	✗	✓	✗	✗	✗
LEGAL	✗	✗	✓	✗	✓	✗	✗
SAFETY	✗	✗	✓	✗	✓	✓	✗

legal consultations, or school communications; disclosure can affect due process or create non-repudiation concerns.

3.2 LINDDUN Framework

LINDDUN [37, 42] is a comprehensive privacy threat modeling framework that supports organizations with identifying and mitigating privacy threats and harms throughout the software development life cycle (SDLC). Each letter of the acronym represents a distinct privacy threat type:

Linking. The ability to connect data across different contexts, revealing patterns or relationships (e.g., linking a user’s health condition mentioned in one conversation to their employer disclosed in another via shared email address).

Identifying. The ability to identify an individual from disclosed information (e.g., combining city, rare disease, and employer to uniquely identify a person).

Non-repudiation. The inability to deny having performed an action, reducing plausible deniability (e.g., signed URLs).

Detecting. The ability to detect the existence of certain data or activities (e.g., API keys reveal active cloud infrastructure projects).

Data Disclosure. Unauthorized or unintended revelation of sensitive data (e.g., health diagnoses, financial account numbers).

Unawareness/Unintervenability. Users’ lack of awareness about data collection or inability to control its use.

Non-compliance. Violations of privacy regulations or policies (e.g., HIPAA, GDPR).

Table 2 operationalizes the mapping between our privacy taxonomy categories and LINDDUN threats, illustrating how the detected PII types translate to concrete privacy harms. For example, CONTACT and PROFILE data primarily pose *Linking*, *Identifying* and *Data Disclosure* threats because the information shared in a chat context would allow the unnecessary identification of an individual.

3.3 Multi-dimensional Risk Scoring Rubric

To evaluate the level of privacy risks, we developed a risk scoring system that operationalizes the assessment over three distinct dimensions, and define an explicit escalation mechanism.

I) Sensitivity: This dimension evaluates the inherent sensitivity of the disclosed information based on the potential harms caused by unauthorized access. Our approach aligns with the CNIL Privacy Impact Assessment (PIA) methodology’s impact severity framework [11], mapping data sensitivity to potential impacts on data

subjects’ rights and freedoms. It expresses these sensitivity levels in a discrete, ordinal scale:

SENSITIVITY SCALE	
Score	Interpretation
Low (1)	Public or non-sensitive information (CNIL: negligible impact)
Moderate (2)	Semi-private data with limited harm potential (CNIL: limited impact—inconvenience, irritation)
High (3)	Confidential information that could cause significant personal or professional harm (CNIL: significant impact—difficulties in exercising rights, fear, reputational damage)
Critical (4)	Highly sensitive data including financial credentials, health records, authentication tokens, or any information involving minors (CNIL: maximal impact—loss of autonomy, financial loss, physical harm)

II) Identifiability: This dimension assesses how directly the disclosed information can be used to identify the specific individual within their conversation context. Again, we define an ordinal scale for this dimension:

IDENTIFIABILITY SCALE	
Score	Interpretation
Quasi-identifier (1)	Demographic or contextual information that contributes to identification when combined with other data points in the same conversation
Common but identifying attribute (2)	Attributes that may be shared by multiple individuals (such as job titles, general location, age ranges) but contribute to identifying a particular individual when disclosed
Context-dependent direct identifier (3)	Information that directly identifies individuals within specific organizational or conversational contexts but lacks universal uniqueness (e.g., employee IDs, organizational roles with specific project details, platform-specific usernames)
Universal direct identifier (4)	Globally-unique identifiers enabling immediate identification across any context (e.g., email addresses, SSNs, biometric data, phone numbers)

III) Exposure: Given that all analyzed conversations are accessible via public share URLs, this dimension focuses on the nature of the information disclosure rather than distribution scope.

EXPOSURE SCALE	
Score	Interpretation
Hypothetical (1)	Mentions of sensitive data categories without exposing actual values
Context only (2)	References to specific sensitive information without revealing actual data
Value present, limited distribution (3)	Actual sensitive values disclosed in conversations with restricted access or limited circulation
Value present (4)	Actual sensitive values disclosed within the publicly accessible conversation context

Note that, since the conversations are publicly available, Score 3 (limited distribution) does not apply to our analysis context.

Aggregated Risk Level Calculation and Escalation Rules:

The overall risk level is determined by evaluating scores across all three dimensions. For base calculation, all dimension scores ≤ 2 yield *Low* risk; any dimension score = 3 results in *Medium* risk; and any dimension score = 4 produces *High* risk. Critical escalation rules (applied only when actual values are disclosed) override the base calculation: findings with Sensitivity = 4 AND Exposure ≥ 3 escalate to *Critical* risk, as do HEALTH or MINOR categories when Exposure ≥ 3 .

Importantly, escalation to *Critical* risk requires both high sensitivity/special category status and actual value disclosure (Exposure ≥ 3). Abstract mentions or hypothetical references (Exposure = 1 or 2) receive lower risk scores regardless of inherent data sensitivity,

as they do not constitute actual privacy breaches. For example: (1) “I need help understanding SSN format” scores Sensitivity=4, Identifiability=4, Exposure=1 \rightarrow *High* risk (not Critical, as no actual SSN disclosed); while (2) “My SSN is 123-45-6789” scores Sensitivity=4, Identifiability=4, Exposure=4 \rightarrow *Critical* risk (actual value in public URL). We illustrate this approach with an example in Appendix C.

4 Methodology

We define a structured processing pipeline, which is depicted in Figure 1 and consists of three distinct phases: (i) Data collection, (ii) Data preparation, and (iii) Evaluation. These are explained in the following sections.

4.1 Data Collection

The main data collection activity has taken place in August 2025. It was implemented by systematically crawling publicly-shared conversation URLs across the OpenAI ChatGPT, Microsoft Copilot, and Google Gemini platforms and led to a consolidated dataset holding in total 102 627 conversations. We systematically collected publicly shared conversation URLs from multiple sources. For ChatGPT, we identified URLs matching the pattern `chatgpt.com/share/` from web archive snapshots and search engine indexes. Similar URL pattern-based approaches were employed for the other two platforms.

In this automated process, each conversation is fetched and encoded in a structured Markdown file. Each conversation is given a unique identifier by adopting the naming convention:

`{platform}_{chat_identifier}_{title}.md`

This chat identifier serves as a unique primary key for subsequent analysis phases.

For ethical reasons, we do not provide further detail about the data collection methodology². This decision protects user privacy and prevents potential misuse of the collection technique, particularly preventing malicious actors from replicating our approach to harvest sensitive user data. It is furthermore important to highlight that the specific shared conversation links collected for this study are in fact no longer accessible as the platforms have since

modified their sharing mechanisms or removed historical shared content in response to public backlash [13]. In addition, the collected conversation datasets will not be made publicly available, rather they will be scheduled for active and definitive deletion after the study. While these fundamental decisions negatively affect the reproducibility and verifiability of our study findings, they have been made to reconcile complex ethical and legal concerns on the collection and processing of these conversations for research purposes. Please refer to Section 7.1 for a detailed discussion of the ethical considerations and related decisions governing this research.

The subsequent two activities –described in the following two sections– were conducted between September and October of 2025.

4.2 Data Preparation

After the collection of chat conversations, this phase transforms raw conversation Markdown into analysis-ready inputs through a tailored filtering and preparation pipeline. We apply three preparatory activities:

User Content Extraction. An AI response might include example PII data (e.g., email addresses) as part of its explanation, which should not be flagged as actual disclosures. For the sake of accuracy in our analysis, we only process user entries within conversations, and we assume that AI responses do not contain any sensitive information³. The user entries extraction process filters conversation blocks to retain exclusively user-generated entries, removing all AI responses, system messages, and metadata.

Snipping conversations. Conversations vary significantly in length, and some exceed the token limits of the LLM used for analysis which requires *snipping the conversation* into smaller chunks. Therefore, for conversations exceeding 20k characters, we employ an entry-aware chunking algorithm that respects user entry boundaries while maintaining analytical continuity. Chunks target 15k characters with configurable overlaps between segments, ensuring that no user entry is fragmented across chunk boundaries (detailed in Appendix B).

Pre-annotation and pseudonymization. We also apply the privacy-enhancing measure of *pre-annotation* before sending conversation text to the LLM for analysis. This approach replaces actual sensitive values with cryptographic placeholders by using specialized regex patterns targeting authentication tokens, financial IDs, contact information, and quasi-identifiers while preserving analytical context through deterministic hash-based keys. Although we utilize a locally hosted LLM for our analysis (as detailed in Section 7.2), this PII scrubbing step serves multiple protective purposes beyond preventing external data exposure. Specifically, pre-annotation prevents potential private data leakage during various research activities, including debugging logs, intermediate file outputs, error messages, and collaborative analysis sessions. Furthermore, this measure protects data subjects throughout the entire analysis pipeline by ensuring that actual sensitive values are never present in our processing infrastructure, even in temporary or cached states. This defense-in-depth approach aligns with privacy-by-design principles and provides additional safeguards against accidental disclosure.

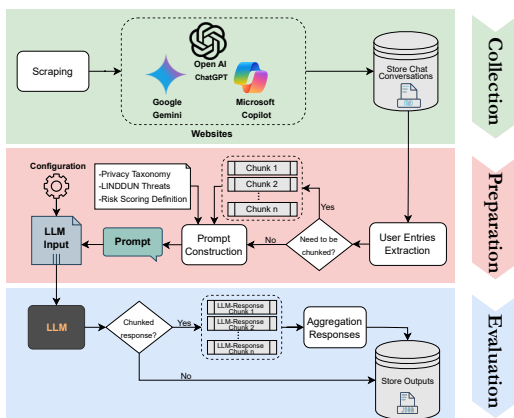


Figure 1: Chat conversation analysis pipeline.

²The detailed approach was checked by the reviewers.

³Or at least, the issue of personal data proliferation in chatbot responses is out of scope of this study.

4.2.1 Prompt Construction. This process is the final preparation step where structured inputs are assembled into comprehensive LLM queries. The analytical prompt combines three main ingredients: (i) conversation snippets, (ii) the privacy taxonomy and threat framework with 44 PII codes from Section 3, and (iii) the risk scoring rubric from Section 3.3. The framework focuses on detecting personal data disclosures; this means that public information (public figures, well-known locations) is not labeled as privacy risks⁴.

The prompt template enforces strict JSON output formatting while embedding conversation metadata. The complete prompt construction details, including the six-layer architecture and structure specifications, are provided in Appendix D. We also present two illustrative examples extracted from our dataset to demonstrate how the privacy risk framework operates in Appendix E.

4.3 Evaluation

The LLM-based analysis was performed using the `gpt-oss:20b` model, a 20-billion parameter open-source language model, deployed locally via Ollama inference server on a dedicated machine equipped with an Intel(R) Xeon(R) Gold 5218R CPU, 64 GB of RAM, and an NVIDIA RTX A5000 GPU with 24 GB of VRAM. LLMs use a *temperature* parameter (0.0-1.0) controlling output variability: higher values increase creativity, while lower values produce more deterministic responses [35]. We configured `temperature=0.1` to maximize reproducibility while preserving minimal variability for complex reasoning. For conversations exceeding 20k characters, we applied automatic chunking targeting 15k-character segments to respect token limitations. Each chunk was processed independently, and the LLM generated structured responses containing detected PII categories, risk assessments, and rationales.

Findings from multiple chunks were aggregated at the conversation level using conservative rules: maximum risk scores across all chunks (e.g., if chunk 1 scored `Sensitivity=2` and chunk 3 scored `Sensitivity=4`, the final score is 4) and union of detected PII categories (all categories identified in any chunk are preserved in the consolidated analysis). All evaluation results were stored in structured JSON format, including: detected PII categories with location hints, risk scores, overall risk classification, LINDDUN threat mappings, and LLM-generated rationales explaining detection logic.

5 Empirical Findings

This section presents a comprehensive statistical analysis of privacy issues identified across the conversations collected over three AI

conversational platforms, addressing research questions **RQ1** and **RQ2** from Section 1. Our analysis involves 100 342 chat conversations from three major platforms: OpenAI ChatGPT (87 035), Google Gemini (10 736), and Microsoft Copilot (2 571). The complete analysis pipeline processed over 218.8 hours of active LLM processing time, achieving a throughput of approximately 310 conversations per hour. We have systematically evaluated these conversations by applying the analytical processing pipeline presented earlier in Figure 1. The statistical findings reveal significant privacy concerns and provide insights into the distribution, severity, and nature of privacy violations in AI-generated content.

Table 3 reports the comprehensive statistics for each platform across five key metrics: initial chat links identified, conversations successfully collected, conversations analyzed through our privacy detection pipeline, total individual privacy issues discovered, and unique conversations containing privacy issues.

The data collection process began with 123 961 total chat links distributed across platforms: ChatGPT (102 683), Gemini (18 493), and Copilot (2 785). The collection phase successfully retrieved a total of 102 627 conversations. Collection failures can be attributed to different root causes. For instance, many ChatGPT share links have become inaccessible over time due to platform policy changes or user privacy settings. They can also be attributed to technical limitations during the scraping process, particularly for large-scale data collection. Regardless, the collection activity had a significant success rate of 82.8%.

The subsequent analysis phase involved processing these collected conversations through our privacy detection pipeline. This phase has led to the successful analysis of 100 342 conversations (97.7% of collected data). Analysis losses or failures occurred due to parsing errors (malformed JSON responses from the LLM evaluation stage or corrupted encoding/special characters in conversation files) and length thresholds (very short conversations below the minimum threshold for meaningful privacy analysis).

The complete analysis resulted in comprehensive privacy assessments across all platforms. Table 3 also reports the number of privacy issues identified per platform. This analysis activity identified a total of 9 623 specific privacy issues distributed across 8 131 unique conversations, indicating that 8.1% of analyzed conversations contain at least one privacy issue. A *privacy issue* refers to any detected instance of private data disclosure that our framework classifies into one or more taxonomy categories (Table 1), maps to LINDDUN threat types, and assigns a risk level based on defined risk scoring rubric in (Section 3.3). This dataset informs the in-depth analysis presented in subsequent sections. For comprehensive conversation length distribution statistics see Appendix B.

Importantly, our findings represent a *lower bound* of privacy risk in AI conversations. Our dataset consists (just) of chats that users considered suitable for public sharing, indicating that private talks (those never meant for external visibility) might contain even more private data disclosures. Thus, the true prevalence of private data disclosures with LLMs is likely higher than our findings indicate. As described earlier, the framework focuses on personal data disclosures, therefore public information such as well-known figures or major locations is not flagged as privacy risk.

Table 3: Data collection and analysis statistics per platform.

Metric	ChatGPT	Gemini	Copilot	Total
Chat links available	102 683	18 493	2 785	123 961
Conversations collected	88 733	11 112	2 782	102 627
Conversations analyzed	87 035	10 736	2 571	100 342
Total privacy issues	7 017	1 618	988	9 623
Conversations with issues	6 135	1 051	945	8 131

⁴Refer to Section 7.2.3 for details on how this capability of our evaluation framework was verified.

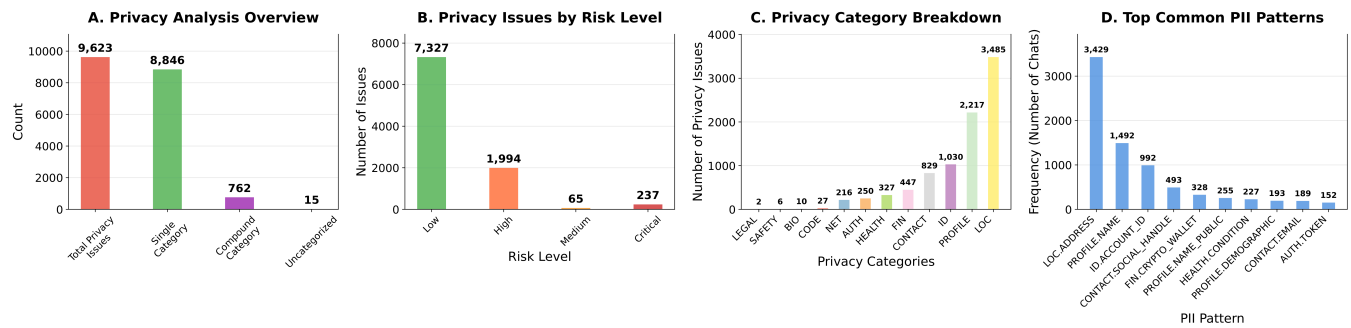


Figure 2: Overview of conversations analyzed and privacy issues detected across all platforms.

5.1 Privacy Issue Distribution

Figure 2 presents a comprehensive synthesis of our privacy analysis performed across 100 342—i.e., the total amount of effectively analyzed conversations—shared chat conversations. The discussion is organized into four interrelated views: A) corpus-level analysis overview, B) risk severity stratification, C) privacy category distribution according to our taxonomy, and D) the most frequently detected PII patterns. Below, we dive more in-depth into the main message as conveyed by each panel of the figure.

Detection and Classification Performance (RQ1). Panel A of Figure 2 demonstrates that our LLM-based pipeline identified 9 623 privacy issues across the analyzed corpus. This corresponds to approximately one privacy issue per ten conversations, revealing the substantial degree of PII disclosure. Of these, 9 608 issues (99.84%) were successfully mapped to our privacy taxonomy (Table 1), with 8 846 issues (91.9%) falling into single categories and 762 issues (7.9%) exhibiting compound categorization patterns spanning multiple privacy domains.

Only 15 identified issues (0.15%) remained uncategorized. The uncategorized subset comprises edge cases resulting from malformed LLM responses (e.g., missing or invalid PII codes in JSON output) and parsing failures where the model generated category labels outside our predefined taxonomy.

Threat Severity (RQ2). The risk-level distribution in Panel B of Figure 2 reveals a critical finding: while 76.1% of detected issues (7 327) are classified as low-risk—typically they involve isolated quasi-identifiers or public-figure references with limited immediate exploitability—, a substantial 23.2% are classified as high-risk (1 984 issues, 20.6%) or critical (247 issues, 2.6%) severity.

High-risk findings predominantly involve direct identifiers such as names, emails, etc., co-occurring with sensitive context (health, financial, or professional disclosures), significantly elevating both identifiability and disclosure harm per the LINDDUN model. Critical-risk cases include disclosed authentication credentials (e.g., API keys, passwords, access tokens), health records with identifiable subjects, financial account details, and data involving minors—each representing imminent exploitation vectors or regulatory non-compliance.

The concentration of 2 231 high-to-critical issues across 100k conversations emphasizes the fundamental issue at hand: privacy breaches via shared chats are not merely an incidental, statistical anomaly but are systematic. These disclosures represent real and

actionable threats to privacy, often arising from user unawareness or underestimation of persistence, of indexability, and technical possibility of cross-platform linkability.

Category-Level Privacy Exposure. Panel C of Figure 2 decomposes the findings by primary taxonomy category, revealing a clear dominance of the Location (3 485 issues, 39.3%) and Profile (2 217 issues, 25%) categories. Together, these two categories account for nearly two-thirds of all detected privacy violations, reflecting users’ tendency to contextualize requests with personal identifiers and geographic anchors (e.g., “I live in [city],” “My address is...”). Location data is particularly high-risk due to its dual role in identifiability (narrowing down individuals via zip code or neighborhood) and detectability (revealing routines, workplace, or home address), with direct implications for physical security (stalking, harassment) and GDPR Art. 9 special-category inference when combined with demographic or health data. The significant presence of the Identity category (ID; 1 030 issues, 11.6%) and Contact information (CONTACT; 829 issues, 9.3%) further amplifies linkability and non-repudiation threats: usernames, email addresses, and phone numbers may serve as persistent join keys across disparate datasets. Financial (FIN; 447 issues, 4%), Health (HEALTH; 327 issues, 3.6%), and Authentication (AUTH; 250 issues, 2.8%) categories, while less frequent, represent potentially disproportionate privacy harms given their sensitivity and regulatory treatment.

Granular PII Pattern Analysis. Panel D of Figure 2 finally drills into the ten most prevalent PII codes, providing further actionable specificity. LOC.ADDRESS dominates with 3 429 detections (of all issues), and it is followed by PROFILE.NAME with 1 492 instances, spanning both self-identifications and third-party mentions; when names co-occur with location or professional details, they enable trivial web-search-based re-identification.

5.2 LINDDUN Threat Analysis

Figure 3 maps the categorized privacy issues to the LINDDUN threat types, showing the prevalence rate of each threat type as a percentage of analyzed conversations per platform. This normalization accounts for differing sample sizes (see conversations analyzed in Table 3) to enable fair cross-platform comparison. The heatmap reveals notable platform-specific differences in threat prevalence. Identifying threats—the ability to single out an individual from a dataset—show varied prevalence rates across platforms, with

Microsoft Copilot exhibiting the highest percentage of conversations containing such threats, followed by Google Gemini, then OpenAI ChatGPT. These threats manifest primarily through direct identifiers (names, emails, phone numbers) and high-cardinality quasi-identifiers (specific addresses, unique usernames, account IDs). Data Disclosure threats, which represent the exposure of sensitive information that could cause harm if revealed, show similar cross-platform patterns. The percentage-based analysis demonstrates that all platforms exhibit substantial prevalence of both Identifying and Data Disclosure threats, confirming that shared conversations routinely contain sufficient signals to re-identify users or third parties mentioned in the dialogue, even without explicit consent or awareness.

This reflects the reality that identifiable data is often simultaneously sensitive (health conditions tied to names, financial details with contact information, professional grievances with employer identification). This coupling amplifies risk: an adversary does not need to perform complex inference attacks when identity and sensitive context are jointly disclosed in plaintext.

6 Analysis of Private Disclosure Patterns

This section addresses research question RQ3 from Section 1 by investigating behavioral patterns in private disclosure through two complementary empirical analyses. Extended conversational interactions present unique privacy challenges where, unlike static text, conversations evolve dynamically. Consequently, users incrementally reveal context, build rapport with AI assistants, and often forget the cumulative disclosure effect as dialogue progresses. This phenomenon may lead users to share increasingly sensitive information in later turns, particularly when initial interactions feel safe and productive.

To investigate this hypothesis systematically, we analyze sensitive disclosure patterns through two complementary lenses:

Conversation Length Analysis (Section 6.1): We examine the relationship between conversation length and privacy risk severity. Our analysis addresses three specific sub-questions: (Q_a) *Do longer conversations exhibit higher rates of privacy issues compared to*

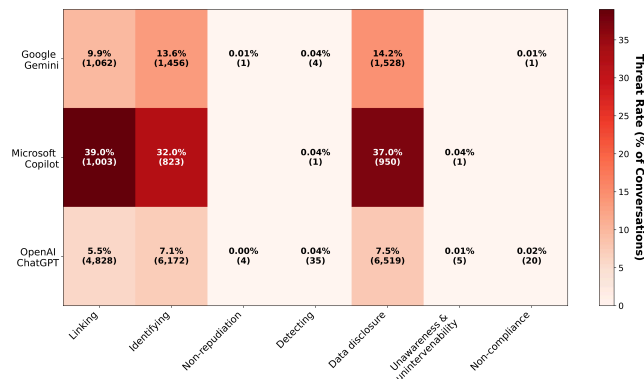


Figure 3: LINDDUN threat analysis across platforms. Values represent the percentage of analyzed conversations exhibiting each threat type, normalized by platform sample size to enable fair cross-platform comparison.

shorter ones? (Q_b) *Is there a relationship between conversation length and the severity of privacy risks encountered?* (Q_c) *Does privacy exposure follow a linear pattern with conversation length, or do certain length thresholds represent inflection points for elevated risk?*

Temporal Disclosure Analysis (Section 6.2): Beyond conversation length, we investigate *when* in the conversation lifecycle private disclosures occur, testing whether users exhibit uniform disclosure patterns or demonstrate temporal concentration in specific conversation phases. The focus of this analysis is on the longer conversations that required chunking due to their substantial length (> 15k characters, see Section 4.2).

Figures 4 and 5 present complementary perspectives on the length-based analysis, while Figure 6 examines temporal distribution patterns across conversation quartiles.

6.1 Conversation Length Analysis

As discussed above (cf., Section 4.2) conversations of more than 15k characters length have been split into conversation chunks as part of the preparation phase. Figure 4 presents a comparative analysis of conversation length (measured in number of chunks) and its relationship to privacy risk. The analysis examines 2 036 multi-chunk conversations (2.03% of the dataset; conversations with more than one chunk), of which 407 (20.0%) privacy issues were identified.

Conversation Length Differs Significantly Between Groups (Addressing Q_a). Panel A (in Figure 4) reveals a striking difference in conversation length between conversations with and without privacy issues. The descriptive statistics show substantial disparities: conversations with privacy issues exhibit a *mean* length of 8.77 chunks, *median* = 4, and *Standard deviation* (Std) = 9.77, while conversations without privacy issues *mean* 4.46 chunks (*median* = 3, *Std* = 4.97).

The difference is highly statistically significant (Mann-Whitney U test⁵, *p* < 0.001) with a medium effect size (Cohen’s *d* = 0.695⁶). This indicates that conversations with privacy issues are, on average, 97% longer than those without issues (mean difference of +4.33 chunks). Extended conversations appear to create more opportunities for disclosure: as users engage in longer interactions,

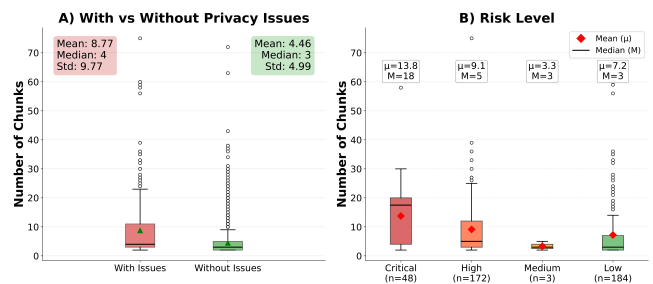


Figure 4: Chunk-based analysis: privacy issues & risk levels.

⁵We employed the Mann-Whitney U test rather than a t-test due to the non-normal distribution of our data, characterized by positive skewness from exceptionally long conversations.

⁶Cohen’s *d* values are interpreted as small (0.2-0.5), medium (0.5-0.8), and large (> 0.8) effects [12]. Our observed effect size of 0.695 indicates a practically meaningful difference beyond mere statistical significance.

they may become more comfortable, provide additional contextual information, and share sensitive data necessary for complex problem-solving or achieving conversational goals.

The wider distribution (higher standard deviation) for conversations with issues also indicates greater variability in disclosure patterns across different conversation lengths.

Risk Severity Correlates with Conversation Length (Addressing Q_b). Panel B (in Figure 4) demonstrates that not all privacy issues are equal—the severity of risk is strongly associated with conversation length. The observations of this panel are:

- **Critical risks emerge in the longest conversations.** Critical-risk conversations are nearly twice as long as the overall average for conversations with issues (13.8 vs 8.77 chunks). The *median* of 18 chunks indicates that high-stakes disclosures typically emerge after extended interaction.
- **High-risk conversations are substantially longer.** High-risk conversations demonstrate the same pattern (9.1 chunks on average) of increased length correlating with increased severity.
- **The *median* tells a nuanced story.** While Critical risks have a *mean* of 13.8 chunks, the *median* is even higher at 18 chunks, indicating that the most common Critical-risk conversations are very long indeed, with a few shorter conversations bringing down the average.
- **Low-risk issues can occur at any length.** Interestingly, low-risk conversations ($n=184$, the largest group) have a *median* of just 3 chunks but a *mean* of 7.2, suggesting that while many low-risk issues occur early, some emerge in longer conversations.

Privacy Risk Increases with Conversation Length (Addressing Q_c). Figure 5 presents a critical finding where the privacy issue rate exhibits a clear positive correlation with conversation length. While our comparative analyses (Section 6) examined all 2036 multi-chunk conversations, for this analysis, the privacy issue rate visualization applied an additional filter to ensure statistical reliability. Specifically, we excluded chunk sizes with fewer than 10 conversations, removing 96 conversations (4.7%) across 40 rare chunk sizes. This conservative threshold is standard practice in rate-based epidemiological visualizations and prevents spurious trends from low-sample-size data points. The resulting dataset of

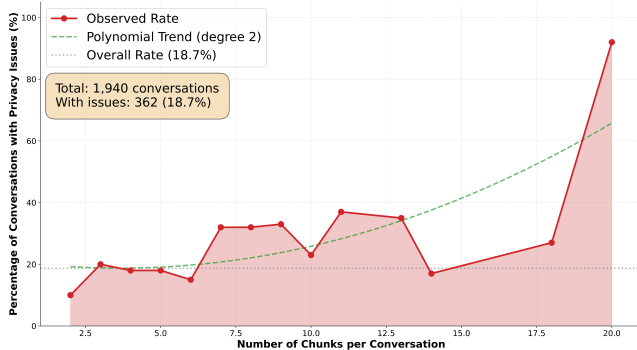


Figure 5: Privacy issue rate as a function of conversation length (1940 out of 2036 multi-chunk conversations; 96 conversations with chunk sizes < 10 excluded for statistical reliability).

1,940 conversations retained 95.3% of the original multi-chunk conversations while providing a statistically robust trend line.

Figure 5 demonstrates that privacy risk escalates non-linearly with conversation length. Conversations exceeding 6 chunks show markedly elevated privacy issue rates (often > 30%), while those beyond 12 chunks frequently approach or exceed 50%. The polynomial trend line reveals an accelerating relationship, where each additional chunk in longer conversations poses disproportionately higher risk than in shorter ones.

6.2 Temporal Analysis of Private Disclosures

To understand *when* in the conversation lifecycle private disclosures occur (**RQ3**), we conducted a temporal analysis of privacy-sensitive information disclosure patterns. In this analysis, we investigate whether users disclose personal information uniformly throughout conversations or exhibit temporal patterns such as early disclosure versus late-stage disclosure, i.e. after several subsequent interactions or conversation turns with the AI chatbot.

Our temporal analysis employs a user-entry- or turn-based quartile approach that tracks the position of private disclosures relative to the conversation’s natural progression. For each finding, we calculate the absolute user entry position, and normalize it by the total number of user entries to determine its relative position within the conversation. This relative position is then mapped to its corresponding conversation quartile: Q1) *Early* [0–25%], Q2) *Early-mid* [25–50%], Q3) *Late-mid* [50–75%], and Q4) *Late* [75–100%].

Figure 6 presents the temporal distribution of private disclosures specifically for multi-chunk conversations (conversations requiring multiple processing chunks due to length). The analysis encompasses 407 multi-chunk conversations containing privacy issues, yielding 1182 privacy findings. This distribution exhibits a strong temporal pattern with marked late-stage concentration: Q1) 149 findings (12.6%), Q2) 143 findings (12.1%), Q3) 181 findings (15.3%), and Q4) 709 findings (60.0%).

A chi-squared test for uniform distribution yields $\chi^2(3) = 774.3$ ($p < 0.001$), providing strong evidence against uniform temporal distribution.⁷ The late-stage concentration is consistent across risk

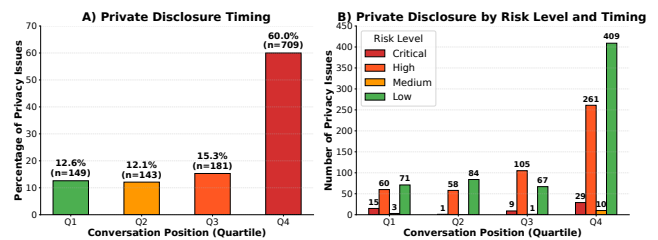


Figure 6: Temporal distribution of private disclosures in multi-chunk conversations.

⁷The chi-square test statistic is calculated as $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$, where O_i represents the observed frequency in quartile i , E_i represents the expected frequency under uniform distribution (295.5 findings per quartile), and $k = 4$ is the number of quartiles. The degrees of freedom is $df = k - 1 = 3$. For our data: $\chi^2 = \frac{(149 - 295.5)^2}{295.5} + \frac{(143 - 295.5)^2}{295.5} + \frac{(181 - 295.5)^2}{295.5} + \frac{(709 - 295.5)^2}{295.5} = 72.6 + 78.7 + 44.4 + 578.6 = 774.3$. The extremely small p-value ($p < 0.001$) indicates that observing such a distribution by chance alone, if the true distribution were uniform, has a probability less than 0.1%.

levels, with even Critical-risk disclosures showing increased prevalence in later conversation stages. Notably, the risk-level breakdown (Panel B) reveals that Low-risk findings (n=631, 53.4%) constitute the largest category, followed by High-risk (n=484, 40.9%), Critical-risk (n=53, 4.5%), and Medium-risk (n=14, 1.2%) disclosures. Despite these different risk profiles, all categories exhibit the characteristic Q4 concentration pattern, with High-risk findings showing particularly strong late-stage clustering (Q4: 261/484, 53.9%).

The temporal analysis shows a striking pattern of late-stage sensitive disclosure, with a **4.8-fold increase** in private disclosure likelihood between the *early* and *late* conversation stages. Furthermore, the quartile distribution reveals relatively consistent disclosure rates across Q1, Q2, and Q3, followed by a dramatic spike in Q4. This pattern indicates a disproportional manifestation of sensitive disclosures in the late conversational phases—particularly in the final quarter of conversations, rather than as a gradual linear increase.

7 Discussion and Threats to Validity

Section 7.1 first discusses the ethical concerns related to empirically evaluating real-world conversation traces. The design of the data processing activity was inspired to the *AoIR* ethical guidelines⁸. Then, Section 7.2 discusses the threats to validity specific to the use of a fully automated LLM-based evaluation approach, and presents our extensive validation and calibration efforts.

7.1 Ethical Considerations

We have performed a crawling activity that involves processing public links to conversations over three distinct platforms. While these conversations have only become publicly accessible *after explicit user action*, the public backlash [8] and the reaction of the chatbot platform providers [13] to make these conversations unavailable indicates, or at very least *suggests*, that this was not well enough understood or known to the affected users. We acknowledge the ethical tension inherent in analyzing these conversation traces collected as such for privacy research: while our work aims to improve user privacy protections, the analysis itself processes contentious private disclosures. We treat all shared conversations as highly confidential, and collect and process this data for the sole purpose of this study.

As we perform observational research and do not engage directly with the data subjects, our primary concern is to maintain confidentiality. To avoid the risks of amplifying public disclosure or repeated exposure in direct consequence of our research activities, we have taken the following key decisions and actions throughout the different stages of our study:

Data Collection. We deliberately refrain from disclosing actionable information about the precise procedures used to obtain links to publicly-available conversation traces, and verified that the issue of public discoverability and enumerability of these links has been addressed in the meantime.

Processing. In the pre-annotation and pseudonymization stage, we actively replace the recognized PII elements with cryptographically-hashed pseudonymous placeholders (e.g., [EMAIL#A7B8C9D2]) before sending these conversation

traces to LLM models for further analysis. These placeholders obfuscate original sensitive values yet can be used for deduplication and maintain in-conversation linkability.

Processing. All raw conversation data is processed locally on secure and isolated infrastructure, with no sensitive content transmitted to external APIs or third party service providers. For this reason, LLM-based analysis was performed only using open-weight models deployed locally.

Processing. With the sole exception of the manual verification of the automated processing pipeline over 200 conversation (as reported in Section 7.2), we have not read or processed the conversations manually, nor have they been distributed within the research team.

Data availability. We will not publish any actual conversation content, neither in raw nor in redacted form.

Reporting. We only report upon aggregated statistics and patterns, and have ensured that there is no residual risk of inference, reconstruction, or reproduction of original data.

Data retention. The conversation trace dataset collected for this study will be irrevocably and definitively deleted after study publication.

Our study (1) increases awareness of privacy risks and implications specific to conversation sharing and publication, and advocates for (2) more proactive in-conversation privacy safeguards, and (3) improved user awareness mechanisms to prevent inadvertent disclosure. We consider the relevancy and importance of this work to outweigh the residual privacy and ethical risks of these research activities.

7.2 LLM-based Processing

Our study uses a processing pipeline based on LLMs, which therefore may be unreliable, inconsistent, or even lead to incorrectness. First, we motivate our choice for applying LLM-based techniques. Then, we discuss our efforts to validate the overall privacy risk assessment framework and processing pipeline used in this study.

7.2.1 Alternative PII detection approaches. Traditional PII detection (e.g., regular expression pattern matching, named entity recognition (NER), rule-based keyword detection, and dictionary-based classification systems) and de-identification techniques (e.g., data masking, anonymization, pseudonymization, suppression, and generalization methods) provide important building blocks but face domain-specific challenges in AI conversations. Rule-based (regex patterns) and dictionary methods excel at recognizing fixed-structure elements (emails, SSNs, credit cards), and are transparent and reproducible. However, they struggle with linguistic and contextual variation, obfuscation, or ambiguous context [23, 32]. Classical and neural NER models improve recall in unstructured text but are optimized for entity typing rather than reasoning about privacy harm; they typically cannot evaluate whether a set of seemingly innocuous mentions (e.g., a city, rare disease, employer) together crosses an identifiability threshold, nor can they weigh exposure scenarios (public URL vs. private workspace) or regulatory implications [27, 31, 32]. Furthermore, domain drift is noticeable in AI chats. The prompts and responses combine code, logs, configuration data, and natural language modalities that differ from the newswire or clinical corpora used to train many NER systems [31, 32].

⁸<https://aoir.org/ethics/>

Table 4: Inter-annotator agreement for manual validation.

Metric	Value	Interpretation
Fleiss' Kappa (κ)	0.799	Substantial
Unanimous Agreement	85.0%	170/200 cases

Recent work shows that LLMs can outperform traditional pipelines for privacy analysis tasks that require contextual reasoning, specifically that LLM-based approaches can lead to superior privacy and utility trade-offs compared to NER approaches [29]. As argued by Źarski et al. [46], conventional NER is ill-suited to assess implications of entity combinations, or to evaluate exposure contexts, which are both central to privacy risk assessment. In our setting, these gaps manifest acutely: AI conversations frequently interleave identifiers with narrative, intent, and sharing metadata (e.g., link visibility), all of which modulate risk.

To address these limitations, we defined a two-stage hybrid methodology that combines deterministic detection precision with the contextual reasoning capabilities of LLMs. Our pipeline therefore deliberately excludes traditional techniques such as NER as the primary detection mechanism.

7.2.2 Framework Validation. To validate the reliability and accuracy of our LLM-based privacy detection framework, we conducted a manual validation study comparing automated LLM classifications against expert human annotations across a stratified random sample of 200 conversations with and without privacy issues.

Validation Methodology. Three independent privacy experts manually reviewed sampled conversations. Each annotator independently assessed whether privacy issues existed, identified privacy categories from the taxonomy, and tagged relevant PII types disclosed. To establish a reliable ground truth, two annotators (R1, R2) underwent a reconciliation process to resolve initial disagreements (21 out of 200 cases, 10.5%) and create a consistent baseline, while a third independent annotator (R3) provided external validation. Majority voting (at least two of three annotators agreeing) determined the final ground truth classification for LLM comparison. Similarly, category and PII assignments required consensus from at least two annotators.

Inter-Annotator Reliability. Human expert agreement provides the baseline for evaluating LLM performance. Table 4 presents inter-annotator agreement statistics. The three annotators achieved a Fleiss' Kappa⁹ of $\kappa = 0.799$, indicating *substantial agreement* approaching near-perfect consensus according to established thresholds [26]. Of the 200 cases, 170 (85.0%) achieved unanimous agreement where all three annotators reached the same classification, while the remaining 30 (15.0%) exhibited disagreement with a 2-1 split, resolved through majority voting.

Following reconciliation between R1 and R2, disagreement arose exclusively from the independent validator (R3), who diverged from the baseline in 30 cases. This 10.0% disagreement rate between the reconciled baseline and independent validator establishes a reference point for human judgment variability in privacy risk assessment, reflecting the inherent subjectivity of such evaluations.

⁹Statistical measure of agreement for categorical data, calculated as $\kappa = \frac{\bar{P} - P_e}{1 - P_e}$, where \bar{P} is the observed agreement proportion and P_e is the expected agreement by chance. Values range from -1 to 1, with higher values indicating stronger agreement.

Table 5: LLM detection performance vs human ground truth.

Precision	Recall	F1	Accuracy	TP	TN	FP	FN
0.83	0.69	0.75	0.78	69	86	14	31

Table 6: Category detection frequency: LLM vs. human.

Category	LLM	R1	R2	R3	Avg Human	Discrepancy
PROFILE	40	76	75	54	68.3	-28.3 (under)
LOC	24	7	4	24	11.7	+12.3 (over)
HEALTH	21	22	20	23	21.7	-0.7
CONTACT	12	11	12	8	10.3	+1.7
ID	6	14	0	7	7.0	-1.0
FIN	5	7	3	6	5.3	-0.3
CODE	3	6	7	0	4.3	-1.3
NET	3	3	6	0	3.0	0.0
AUTH	1	4	3	1	2.7	-1.7
Total	115	150	130	123	134.3	-19.3 (under)

LLM Performance Metrics. Comparing LLM-generated classifications against the majority-vote ground truth yielded the performance metrics shown in Table 5. The framework achieved 77.5% overall accuracy, with precision of 83.1% (proportion of correctly identified positives) and recall of 69.0% (proportion of actual positives detected), resulting in an F1-score of 0.754. The classification breakdown reveals 69 true positives, 86 true negatives, 14 false positives (over-detection), and 31 false negatives (under-detection).

Contextualizing these results against the human baseline performance reveals that the LLM disagreement rate is approximately twice that observed between the reconciled annotators and the independent validator. The LLM disagreed with the majority-vote ground truth in 22.5% (45 out of 200 cases), compared to the 10.0% disagreement rate between the reconciled baseline (R1/R2) and the independent validator (R3). While the LLM exhibits higher disagreement, this pattern remains within a bounded range of human expert variability, suggesting that the framework's deviations represent judgment variations.

Category-Level Detection Patterns. Table 6 presents detection frequencies across privacy categories for the LLM and human annotators. Two systematic discrepancies emerge: the LLM under-detects PROFILE category violations (40 vs. 68.3 human average) while over-detecting LOC category issues (24 vs. 11.7 human average). These patterns align with the false negative and false positive distributions observed in our error analysis—PROFILE under-detection stems primarily from missed professional context (job titles, employers) requiring multi-hop inference, while LOC over-detection reflects conservative flagging of generic geographic references (e.g., city names) that human annotators correctly dismissed as non-identifiable. Overall, the LLM identified 115 total category instances compared to an average of 134.3 by human annotators, indicating a net under-detection tendency of 14.4% across all privacy categories. This conservative bias, while reducing false positives in individual categories, results in lower recall for context-dependent privacy risks that require deeper semantic understanding.

7.2.3 Distinguishing Personal Disclosure from Public Information.

Our study relies on the LLM's contextual reasoning capability to correctly distinguish disclosures pertaining public information

from those pertaining personal data. This means that for example, references to public figures (e.g., Donald Trump) will not as such be flagged as privacy risks.

To validate this capability and confirm that the framework is sensitive to private disclosure, we have defined a number of synthetic test cases. These test cases specifically verify and demonstrate the capability to correctly flag personal disclosures, yet also to filter out non-personal references to public information. Detailed test results and analysis are provided in Appendix F.

7.3 Theoretical Interpretations of Private Data Disclosure Patterns

Our RQ3 results show strong correlations between conversation length and private data disclosure, non-linear escalation of disclosure severity, and higher concentrations of private data disclosures in later conversation stages. These results align with established theories of human relationship formation:

Social Penetration Theory (SPT) [7] states that interpersonal relationships develop through gradual self-disclosure, progressing from superficial exchanges to deeper revelations as trust builds through mutual vulnerability. Recent research extends SPT to human-AI interactions, demonstrating that users develop similar patterns of progressive disclosure with conversational agents [38]. Applying SPT to our study results, the observed patterns become interpretable as normative relationship-building behaviors: extended conversations create more opportunities for disclosure as users become comfortable; the non-linear escalation in risk severity reflects progression from low-stakes to high-stakes revelations as perceived intimacy deepens; and the late-stage temporal concentration may indicate critical decision-seeking moments where users seek advice on sensitive matters after establishing perceived rapport.

However, these patterns could equally reflect task structure (users provide background context early, specific details later), problem-solving dynamics (solutions require sensitive information), or topic complexity effects rather than intimacy mechanisms.

Interpersonal Process Model (IPM) [34] frames intimacy as interactive feedback loops: sharing information elicits responsive feedback, which validates and promotes further sharing. Applying this model to human-AI interactions suggests that chatbot responsiveness through perceived understanding, relevance, and personalization may create a feedback loop that progressively lowers disclosure barriers. The cumulative effect across longer conversations may build perceived relational depth, explaining why extended interactions correlate with higher-severity private data disclosures.

Here also, alternative interpretations remain plausible: longer conversations may simply involve more complex topics requiring detailed context, independent of any intimacy-building mechanism. The correlation between length and severity could reflect problem complexity rather than relational dynamics.

While our study results are **not inconsistent** with these theoretical interpretations, they are not directly validated by our exploratory study. The above theoretical interpretation is performed *post-hoc*, in an attempt to explain the identified patterns.

Further research is required to properly distinguish between competing explanations. Such validation would require direct user studies with controlled experiments manipulating conversational

features to test causal pathways, think-aloud protocols during sharing decisions to understand in-the-moment reasoning, and longitudinal interviews exploring disclosure motivations.

8 Related Work

We first discuss work aimed to understand user behaviors and expectations on privacy in conversational chatbots. Then, we discuss the technical privacy evaluation of chatbot applications.

Privacy Attitudes and Expectations of Chatbot Users. Belen Saglam et al. [9] surveyed 491 British citizens to understand their attitudes and perceptions on personal data disclosures to chatbots. The study identifies common concerns about data retention and loss of control over information collected in online automated conversations. While the degree of concern was shown to be unaffected by gender or education level, it was positively correlated to age.

The study also pinpoints the grammatical correctness and overall quality of chatbot responses as primary factors influencing user trust. A similar finding is shared by Ischen et al. [19] and Hong et al. [18] who show that anthropomorphic, human-like cues in chatbot responses and behaviors lead to higher sensitive disclosures.

Kwesi et al. [25] have conducted a semi-structured interview study with 21 users of mental health support chatbots. They find that the capability of these services to generate natural language and perform follow-up conversation creates a false sense of trustworthiness among users, and define ‘*intangible vulnerability*’ in reference to users undervaluing and underestimating the privacy impacts of sharing of mental health information. While users do recognize that they willingly share highly sensitive information, they fail to concretize the longer-term privacy harms.

Tran et al. [41] have conducted a survey study among 300 ChatGPT users to understand privacy norms around data sharing. Here, over 80% of respondents indicated that they view chat data as sensitive. The study also shows that users have strong concerns about privacy and the secondary use of chatbot data, especially when data is combined from different contexts.

Grosse et al. [14] have surveyed 3 270 UK adults about their behavior in online conversational chats, seeking to understand the extent to which users of conversational chatbots exhibit specific risk behaviors (such as disclosing sensitive information, e.g., passwords), and about their awareness of the option of opting out of any secondary use of this data supplied by them. The study highlights the overall lack of guardrails to prevent such behavior or in the very least, to properly inform these users of the extensive security and privacy risks they incur through such behavior.

Zhan et al. [47] have conducted a randomized-controlled trial with 502 participants that were exposed to a malicious chatbot, i.e. an LLM that implements different strategies to encourage and provoke disclosures of personal information from users. This study shows that some tactics are very effective, i.e. they successfully lead to users disclosing personal information without triggering any concern or awareness, or without causing any discomfort among participants. This study also emphasizes a key challenge: ‘*LLM models may make [inferences] beyond explicit personal information.*’. Even in cases when the shared information is not complete or directly identifiable, the LLM may inadvertently perform non-trivial inference of personal information.

Chung et al. [10] have studied the phenomenon of *privacy fatigue* specifically among ChatGPT users. They have surveyed 695 undergraduate students in South Korea, and highlight an alarming sentiment of hopelessness, weariness, and complacency regarding privacy among these respondents.

Zufferey et al. [49], conducted a complementary quantitative survey study with chatbot users (ChatGPT). They examined privacy risks from a user-centered perspective on 422 participants by comparing behaviors and concerns across three groups: service users, local users (i.e., users of a local instance of a conversation service), and non-users. Their findings reveal users to generally exhibit lower privacy concerns than non-users, yet frequently share sensitive data related to lifestyle, health, standard of living, and opinions while rarely disclosing personal identifiers or credentials. Notably, partially local users (employing both external and local services) tend to share more data and use them for more diverse purposes. The study highlights a privacy paradox where increased concern does not translate to reduced sharing of sensitive information; approximately half of users indicate willingness to stop using conversational chatbot for privacy reasons such as data leaks.

A significant body of work adopts a user survey-based methodology to understand user norms, attitudes, and behaviors in the use of AI chatbots. However, there is a ‘privacy paradox’, a clear discrepancy between self-reported privacy attitudes, and the actual privacy behaviors of users [21, 49]. This is to our knowledge the first study to evaluate privacy implications of AI chatbots over a large-scale, and real-world conversation dataset.

Privacy Analysis of AI Chatbot Applications. Ragab et al. [33] performed static and dynamic analysis over 21 distinct Android romantic chatbot apps to evaluate the degree of compliance between the conversational behavior of these apps and their privacy policies. This study unearths a range of problematic issues, including a general lack of compliance to the privacy policy, lacking customer service, inadequate age verification, and extensive usage of tracking services. The extensive use of tracking services in this context is worrisome as this involves the systemic sharing and sensitive disclosure to third parties.

Yener et al. [45] evaluated the privacy settings, in-app privacy controls, and privacy policies of 12 healthcare AI chatbots, and find the majority of these apps to be lacking on these aspects.

The literature review study of Silva et al. [36] has led to eight fundamental privacy requirements for chatbot applications. Some of these key requirements –such as (Req1) automated recognition and protection of personal data shared in a chat context, and (Req3) protecting access to conversation histories– are strongly violated by the publication and sharing of conversation histories.

Spearheaded in the domain of processing medical text, sensitive disclosure through unstructured textual information has been an active research topic for over two decades [16, 24, 28]. Recent NLP innovations [22, 43] such as sentence transformers have been shown effective as an alternative to traditional named entity recognition (NER) approaches. Additionally, some efforts have started using general-purpose LLMs for evaluation and recognition of sensitive disclosure in text [20, 44]. Misregallah et al. [31] have used the Text Analytics AI service of Azure to identify PII over a sampled

selection of conversations taken from the WildChat dataset [48]. Their conclusion is that traditional PII detection is insufficient for evaluating sensitive disclosure in chatbot conversations.

In this paper, a novel, two-phased LLM-based methodology is used to evaluate textual conversation traces for sensitive disclosure. This comprehensive methodology is aligned with relevant privacy frameworks, draws from established privacy evaluation methods such as LNDDUN, and uses LLM prompting techniques such as knowledge pre-prompting and step-by-step reasoning. The methodology has been validated against four open-weight LLMs, and has been successfully applied at large scale.

9 Conclusion

This paper presents a comprehensive empirical analysis of private disclosures and their ensuing privacy risks in AI chatbot conversations, based on over 100,000 real-world shared dialogues across OpenAI ChatGPT, Microsoft Copilot, and Google Gemini. The analysis leverages a novel hybrid privacy detection pipeline that combines deterministic pre-annotation with LLM-based contextual reasoning. Our findings reveal that roughly 8% of shared conversations contain privacy-sensitive information, including personal identifiers and health-related details. Notably, nearly one-quarter of these cases involve high-risk or critical-severity exposures that pose immediate threats to user privacy. Furthermore, extended conversations exhibited 2.5 times higher privacy risk rates compared to brief exchanges, with the most severe risks emerging almost exclusively in lengthy dialogues. We argue that these results represent a lower bound, as our automated framework analyzed only written user inputs, and excluded AI-generated responses or uploaded files. The manual validation efforts revealed several cases where users uploaded confidential files (e.g., CVs, medical prescriptions), and in other cases, we observed that the private disclosures stemmed from the chatbot response (e.g., because it remembered earlier conversations).

Our results demonstrate that the issues and risks of private disclosure in AI chatbot conversations are real and systemic, and increase with conversation length. Addressing these risks requires a multi-faceted approach: first, AI chatbot platform designers must tackle the persistent vulnerabilities inherent in sharing mechanisms, recognizing that de-indexing or deletion alone cannot fully eliminate risks once conversations are archived. Second, systems should anticipate and manage the progressive nature of user disclosure over time. Third, our findings call for a new class of privacy-enhancing techniques specifically designed to assist users in avoiding inadvertent personal information disclosure during interaction, complementing existing guardrails for safety and classical privacy-enhancing techniques (e.g., differential privacy) that primarily address training data leakage. Ultimately, mitigating these risks will require substantial and coordinated efforts among platform developers, policymakers, and researchers.

Artifact availability. The privacy risk assessment framework used in this study is fully available in this repository¹⁰, together with reproducible validation test cases.

¹⁰<https://github.com/Majid-Mollaefar/chatbot-confessions-artifact>

Acknowledgments

This research is partially funded by the Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007) and the European Union (Next Generation EU), by the Research Fund KU Leuven, and by the Cybersecurity Research Program Flanders.

Grammarly was used for grammar and spelling verification, and ChatGPT to improve the clarity of complex sentences.

References

- [1] 2003. HIPAA Privacy and Security Rules, 45 CFR Parts 160 and 164.
- [2] 2011. ISO/IEC 29100: Information technology – Security techniques – Privacy framework.
- [3] 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union.
- [4] 2019. ISO/IEC 27701:2019 – Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management.
- [5] 2020. *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management*. Technical Report. NIST.
- [6] 2022. Payment Card Industry Data Security Standard, Version 4.0.
- [7] Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston.
- [8] Ashley Belanger. 2025. ChatGPT users shocked to learn their chats were in Google search results. <https://arstechnica.com/tech-policy/2025/08/chatgpt-users-shocked-to-learn-their-chats-were-in-google-search-results/>.
- [9] Rahime Belen Saglam, Jason RC Nurse, and Duncan Hodges. 2021. Privacy concerns in chatbot interactions: When to trust and when to worry. In *International Conference on Human-Computer Interaction*. Springer, 391–399.
- [10] Jiwon Chung and Hun-yeong Kwon. 2025. Privacy fatigue and its effects on ChatGPT acceptance among undergraduate students: is privacy dead? *Education and Information Technologies* (2025), 1–23.
- [11] CNIL. 2018. Privacy Impact Assessment (PIA) – Methodology. <https://www.cnil.fr/en/PIA-privacy-impact-assessment-en> Section 1.1: Assessing the severity of potential impacts.
- [12] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.
- [13] Vlad Contantinescu. 2025. Your Shared ChatGPT Chats May Be Publicly Searchable-Here's How to Delete Them. Bitdefender. <https://www.bitdefender.com/en-us/blog/hotforsecurity/your-shared-chatgpt-chats-may-be-publicly-searchable-heres-how-to-delete-them>
- [14] Kathrin Grosse and Nico Ebert. 2025. Prevalence of Security and Privacy Risk-Inducing Usage of AI-based Conversational Agents. *arXiv:2510.27275 [cs.CR]* <https://arxiv.org/abs/2510.27275>
- [15] Ece Gumusel. 2025. A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology* 76, 1 (2025), 121–154.
- [16] Guntur Budi Herwanto, Gerald Quirchmayr, and A Min Tjoa. 2021. A named entity recognition based approach for privacy requirements engineering. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 406–411.
- [17] Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. *Context Rot: How Increasing Input Tokens Impacts LLM Performance*. Technical Report. Chroma. <https://research.trychroma.com/context-rot>
- [18] Yoonsuh Hong, Xinge Li, and Yongjun Sung. 2025. How anthropomorphism impacts users' self-disclosure and evaluation of empathetic conversational agents. *Interacting with Computers* (09 2025), iwaf042. <https://doi.org/10.1093/iwc/iwaf042> *arXiv:https://academic.oup.com/iwc/advance-article-pdf/doi/10.1093/iwc/iwaf042/64286757/iwaf042.pdf*
- [19] Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda Van Noort, and Edith Smit. 2019. Privacy concerns in chatbot interactions. In *International workshop on chatbot research and design*. Springer, 34–48.
- [20] Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang. 2024. Llm based biological named entity recognition from scientific literature. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 433–435.
- [21] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security* 64 (2017), 122–134. <https://doi.org/10.1016/j.cose.2015.07.002>
- [22] Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. 2024. From zero to hero: Harnessing transformers for biomedical named entity recognition in zero-and few-shot contexts. *Artificial Intelligence in Medicine* 156 (2024), 102970.
- [23] Poornima Kulkarni and NK Cauvery. 2021. Personally identifiable information (pii) detection in the unstructured large text corpus using natural language processing and unsupervised learning technique. *International Journal of Advanced Computer Science and Applications* 12, 9 (2021).
- [24] Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1937–1945.
- [25] Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. 2025. 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)*. 6007–6024.
- [26] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
- [27] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 346–363.
- [28] Bradley Malin and Latanya Sweeney. 2004. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics* 37, 3 (2004), 179–192.
- [29] Benet Manzanara-Salor and David Sánchez. 2025. A comparative analysis, enhancement and evaluation of text anonymization with pre-trained Large Language Models. *Expert Systems with Applications* (2025), 129474.
- [30] Erika McCallister, Timothy Grance, and Karen Scarfone. 2010. *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. Technical Report SP 800-122. NIST.
- [31] Niloofar Mirehghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-LLM conversations in the wild. *arXiv preprint arXiv:2407.11438* (2024).
- [32] Karthik Murugadoss, Ajit Rajasekharan, Bradley Malin, Vineet Agarwal, Sairam Bade, Jeff R Anderson, Jason L Ross, William A Faubion, John D Halamka, Venky Sundararajan, et al. 2021. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns* 2, 6 (2021).
- [33] Abdelrahman Ragab, Mohammad Mannan, and Amr Youssef. 2024. "Trust Me Over My Privacy Policy": Privacy Discrepancies in Romantic AI Chatbot Apps. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 484–495.
- [34] Harry T. Reis and Phillip Shaver. 1988. *Intimacy as an interpersonal process*. USA.
- [35] Krishna Ronanki, Beatriz Cabrero-Daniel, Jennifer Horkoff, and Christian Berger. 2024. Requirements engineering using generative AI: Prompts and prompting patterns. In *Generative AI for effective software development*. Springer, 109–127.
- [36] Geovana Ramos Sousa Silva and Edna Dias Canedo. 2025. Privacy in Chatbot Conversation-Driven Development: A Comprehensive Review and Requirements Proposal. *ACM Transactions on Software Engineering and Methodology* (2025).
- [37] Laurens Sion, Dimitri Van Landuyt, Kim Wuyts, and Wouter Joosen. 2025. Robust and reusable LINDDUN privacy threat knowledge. *Computers & Security* 154 (2025), 104419.
- [38] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.
- [39] Chris Stokel-Walker. 2025. Google is indexing ChatGPT conversations, potentially exposing sensitive user data. Fast Company. <https://www.fastcompany.com/91376687/google-indexing-chatgpt-conversations>
- [40] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
- [41] Sarah Tran, Hongfan Lu, Isaac Slaughter, Bernease Herman, Aayushi Dangol, Yue Fu, Lufei Chen, Biniyam Gebreyohannes, Bill Howe, Alexis Hiniker, et al. 2025. Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2522–2534.
- [42] Dimitri Van Landuyt, Laurens Sion, and Wouter Joosen. 2025. LINDDUN MAE-STRO: an architecture framework for privacy threat modeling. *Software And Systems Modeling* (2025), 1–37.
- [43] Weijie Wang, Xiaoying Li, Huiling Ren, Dongping Gao, and An Fang. 2023. Chinese clinical named entity recognition from electronic medical records based on multiseamantic features by using robustly optimized bidirectional encoder representation from transformers pretraining approach whole word masking and convolutional neural networks: model development and validation. *JMIR Medical Informatics* 11, 1 (2023), e44597.
- [44] Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568* (2024).
- [45] Ramazan Yener, Guan-Hung Chen, Ece Gumusel, and Masooda Bashir. 2025. Can I Trust This Chatbot? Assessing User Privacy in AI-Healthcare Chatbot Applications. *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 809–820.
- [46] Tomasz Leszek Żarski and Artur Janicki. 2025. Enhancing Privacy While Preserving Context in Text Transformations by Large Language Models. *Information* 16,

- 1 (2025), 49. <https://doi.org/10.3390/info16010049>
- [47] Xiao Zhan, Juan Carlos Carrillo, William Seymour, and Jose Such. 2025. Malicious LLM-Based Conversational AI Makes Users Reveal Personal Information. arXiv:2506.11680 [cs.CY] <https://arxiv.org/abs/2506.11680>
- [48] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470* (2024).
- [49] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. 2025. "AI is from the devil." Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies* 2025, 3 (2025), 5–28.

A Alignment between Privacy Taxonomy and Normative Frameworks and Regulations

Table 7 presents a mapping and anchoring of the core taxonomy classes to the various regulatory and normative frameworks [1–6, 30]. This mapping supports the translation and contextualization into compliance risks.

B Conversation Length Distribution

To provide context for our analysis methodology and the chunking threshold decisions described in Section 4.2, Table 8 presents some statistics for conversation lengths (measured in characters) across our dataset. Note that these statistics consider only user-provided text, excluding assistant responses.

The distribution reveals substantial variation in conversation lengths. The typical conversation is quite compact, with a median of 335 characters (roughly a few sentences), while the mean of 2 626 characters is pulled upward by longer conversations. The high standard deviation (15 791 characters) and extreme maximum (941 243 characters) demonstrate the distribution’s heavy right skew.

The data reveals that only 2.1% of conversations exceed 20k characters, meaning the vast majority (97.9%) are analyzed as complete, unbroken units. This distribution naturally supports whole-conversation analysis for most cases while requiring a chunking strategy for exceptionally long dialogues.

During initial analysis without chunking, we observed that the LLM occasionally yielded no response for very long conversations. Recent research on long-context LLM performance [17] demonstrates that model reliability degrades non-uniformly as input length increases—a phenomenon termed “*context rot*”—where performance becomes increasingly unreliable with longer inputs even on simple tasks. Given these empirical observations, we adopted a chunking strategy with 15k characters as the target chunk size in order to have (i) sufficient context (5.7 times median length), and (ii) fit within the model’s context window alongside prompt overhead.

Note that these thresholds are configurable parameters in our methodology rather than fixed constraints, allowing adaptation to different LLM architectures with varying context window capacities.

C Risk Assessment Framework Illustration

This section provides a complete worked example illustrating the multi-dimensional risk assessment framework described in Section 3.3.

User Chat Conversation

```
### USER ENTRY 1 ###
Hi! I'm having trouble setting up my development
environment. Can you help me configure my email settings?
### USER ENTRY 2 ###
My work email is sarah.johnson@techcorp.com and I need
to configure it with the SMTP server. The authentication
keeps failing.
### USER ENTRY 3 ###
I'm a senior software engineer at TechCorp working on the
authentication module. This is really blocking my progress
on the project.
```

After the Data preparation step, the annotated version (privacy preserved), produced in case of using a non-local LLM, will appear as follows:

Pre-annotation Processing

```
### USER ENTRY 1 ###
Hi! I'm having trouble setting up my development
environment. Can you help me configure my email settings?
### USER ENTRY 2 ###
My work email is [EMAIL#A7B8C9D2] and I need to configure
it with the SMTP server. The authentication keeps failing.
### USER ENTRY 3 ###
I'm a [EMPLOYER#F3E4D5C6] software engineer at
[EMPLOYER#F3E4D5C6] working on the authentication module.
This is really blocking my progress on the project.
```

In this example, we have a disclosure of email address; therefore we assign a Sensitivity score of 2 (work email has moderate sensitivity with professional impact). We attribute Identifiability score of 4 (direct, unique identifier), and Exposure score of 4 (actual value disclosed in a public/shared context). The resulting Risk Level is *High* because a dimension score of 4 elevates the finding to High even if other scores are lower. The structured output is illustrated in the following box.

JSON schema output

```
{
  "chat_id": "example_01",
  "chunk_index": 1,
  "findings": [{
    "pii_code": "CONTACT.EMAIL",
    "subject_type": "private_individual",
    "evidence": {
      "evidence_type": "value_present",
      "location_hint": "user entry #2 in this chunk",
      "risk": {
        "sensitivity": 2,
        "identifiability": 4,
        "exposure": 4,
        "risk_level": "High",
        "threats": ["Identifiability", "Disclosure"],
        "regulatory": ["GDPR"],
        "normalized_key": "[EMAIL#A7B8C9D2]"
      }
    }
  }]
}
```

D Prompt Construction

The construction of an analytical prompt, as shown in Figure 7, is based on combining three main ingredients:

- (1) Conversation snippets (single/multiple chunk(s)), as described in Section 4.2.

Table 7: Mapping between the taxonomy classes and the normative regulatory frameworks.

Category	PII codes
CONTACT	NIST SP 800-122 specifically lists direct identifiers such as name, address, email, and telephone number among canonical examples; GDPR Art. 4(1) and Recital 30 cover names and online/contact identifiers as personal data; ISO/IEC 29100 defines PII and principles governing collection/use of contact details, and ISO/IEC 27701 provides PIMS controls and GDPR mappings for handling such PII; the NIST Privacy Framework’s Identify/Control functions call for inventorying and minimizing common PII including contact data. In health contexts, HIPAA’s de-identification Safe Harbor (45 CFR 164.514(b)(2)(i)) enumerates names, geographic subdivisions smaller than a state (addresses), telephone/fax numbers, and email addresses as identifiers requiring removal.
PROFILE	Re-identification of individuals is considered in NIST SP 800-112 [30] and in GDPR [3]. ISO/IEC 29100 and 27701 provide privacy governance for collection limitation and purpose specification.
ID	NIST SP 800-122 explicitly lists SSN, driver’s license numbers, and passport numbers among direct identifiers requiring the highest confidentiality protections; GDPR Art. 4(1) and Recital 30 recognize national identification numbers as personal data with heightened sensitivity; the NIST Privacy Framework calls for minimizing collection and use of high-risk identifiers; HIPAA’s Safe Harbor de-identification standard (45 CFR 164.514(b)(2)(i)) requires removal of SSN, medical record numbers, health plan beneficiary numbers, and account numbers.
BIO	Biometric identifiers used for uniquely identifying a person (e.g., fingerprints, faceprints, etc.) are treated as special-category data under GDPR and as sensitive PII in privacy frameworks [2, 3]. ISO/IEC 27701 supports documenting lawful bases and minimization controls for such processing [4].
AUTH	NIST SP 800-122 frames authentication information linked to individuals as PII requiring confidentiality safeguards; GDPR Art. 5(1)(f) and Art. 32 impose integrity and security of processing, making credential protection a direct obligation even when secrets merely enable access to personal data; ISO/IEC 29100’s <i>security safeguards</i> principle and ISO/IEC 27701’s PIMS controls require access control, minimization, and documented measures for protecting PII and related authentication data; the NIST Privacy Framework’s Identify/Control functions call for integrating security measures to prevent unauthorized access to data, explicitly encompassing credential handling; PCI DSS prescribes strong authentication and secret handling within the cardholder data environment (e.g., requirements on credential management, MFA, and protection of sensitive authentication data); HIPAA’s Security Rule mandates technical safeguards, including unique user identification and person/entity authentication (45 CFR 164.312).
NET	Technical identifiers such as IP-addresses are recognized under GDPR; NIST SP 800-122 recognizes technical identifiers as PII in certain contexts. Governance anchors include ISO/IEC 27701 controls and NIST Privacy Framework outcomes that integrate privacy with security monitoring.
FIN	PCI DSS defines the scope for cardholder data (CHD) and sensitive authentication data (SAD), prescribing controls such as encryption, masking, key management, access control, and monitoring; GDPR treats financial identifiers as personal data when linked to an individual and requires appropriate security of processing (Art. 32), with Recital 75 highlighting risks of fraud and identity theft; NIST SP 800-122 lists credit card and bank account numbers among canonical PII examples requiring confidentiality protections; ISO/IEC 29100 defines PII and governing principles applicable to financial details, while ISO/IEC 27701 provides PIMS controls and mappings to regulatory expectations for handling such data; the NIST Privacy Framework calls for inventorying, minimizing, and protecting high-risk PII, including financial identifiers.
HEALTH	GDPR classifies health data within the special category (Art. 9) requiring additional protections; HIPAA defines Protected Health Information (PHI) in covered healthcare contexts. ISO/IEC 27701 and ISO/IEC 29100 provide governance anchors for handling sensitive categories.
LOC	GDPR recognizes “online identifiers” and location data as personal data when it can relate to an identifiable person; ISO/IEC 29100 addresses context and proportionality in processing such data.
LEGAL/SAFETY	GDPR’s principles (lawfulness, fairness, confidentiality) apply when legal identifiers can be linked to individuals.

Table 8: Conversation length statistics.

Metric	Value	Interpretation
Total conversations	100,342 [*]	Dataset size
Minimum length	22	Shortest conversation
Maximum length	941,243	Longest conversation
Mean length	2,626	Average length
Median length	335	Typical length
Standard deviation	15,791	Variability in lengths
Conversations > 15k chars	2,637 (2.63%)	Exceed target chunk size
Conversations > 20k chars	2,102 (2.09%)	Require chunking

^{*}Five conversations contained no user text and were excluded from length statistics.

- (2) The privacy taxonomy and threat framework defined in Section 3. To have a systematic privacy risk assessment, we have integrated the PII taxonomy comprising 44 codes (see Table 1) across categories defined in Section 3.1.
- (3) The risk scoring rubric defined in Section 3.3, which provides specific instructions for assessing privacy risks along three dimensions (sensitivity, identifiability, exposure) with explicit escalation rules.

We also add LINDDUN threats as part of the knowledge base for the LLM to map identified PII categories to relevant privacy threats.

The prompt template enforces strict JSON output formatting while embedding conversation metadata, including unique chat identifiers, chunk positioning information, and user entry delimiters that maintain positional context during multi-chunk analyses.

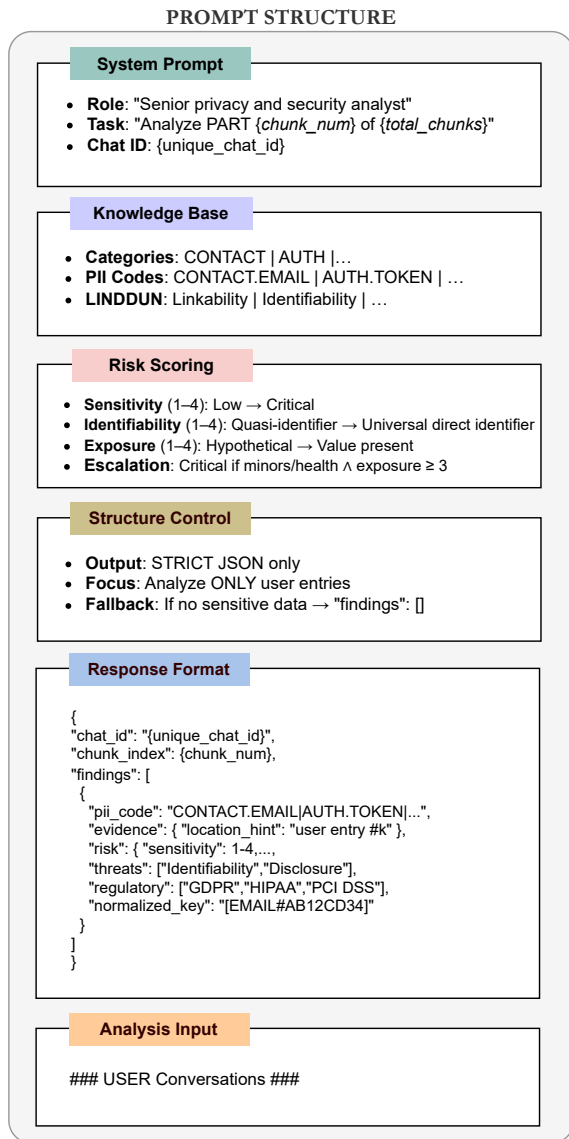


Figure 7: Overview of the prompt structure of our analysis pipeline.

Figure 7 presents the prompt construction as a six-layer architecture comprising:

- i) *System Prompt*—a concise description of the assistant’s role (“Senior privacy and security analyst”), the analysis task for a given chunk of user entries content, and the chat identifier;
- ii) *Knowledge Base*—the taxonomy categories (Section 3.1) with their PII codes (Table 1) and the LINDDUN threat classes;
- iii) *Risk Scoring Instructions*—the three-dimensional rubric from Section 3.3 over sensitivity, identifiability, and exposure with explicit escalation rules for special categories (e.g., minors and health);

- iv) *Structure Control*—hard constraints enforcing JSON-only output, placeholder-based normalization, and empty findings when no sensitive data is detected;
- v) *Response Format*—a fixed JSON schema guaranteeing structured outputs;
- vi) *Analysis Input*—the user conversation to be analyzed.

The constructed prompt, combined with fixed configuration parameters (e.g., model, temperature, and maximum token count), forms the input to the LLM.

E Illustrative Examples

This section presents representative conversation examples extracted from our dataset demonstrating how our LLM-based privacy detection system operates in practice. We selected two examples that illustrate the system’s detection and scoring capabilities.

Example 1: Health Information Disclosure. In this conversation, the user seeks medical advice from the AI assistant by sharing comprehensive personal and health-related information. The LLM-based evaluation detected this conversation with a *critical* risk level, representing one of the highest privacy risk scenarios identified in our dataset. The LLM model assigned these values *Severity=4*, *Identifiability=1*, and *Exposure=4* to three risk scoring dimensions. It identified these privacy categories impacted: HEALTH.CONDITION, HEALTH.MEDICATION, PROFILE.DEMOGRAPHIC. The associated LINDDUN threats are “*Identifiability*” and “*Disclosure*”, with the rationale that *health information is highly sensitive; disclosure can lead to discrimination or stigmatization*.

Conversation Transcript 1

◀ **USER:**

I’m a 50-year old man that weighs 180 pounds. I take medication for high cholesterol and high blood pressure. I’m having heartburn, nausea, and shoulder pain. What might be wrong with me?

▶ **ASSISTANT:**

I’m sorry to hear that you’re experiencing these symptoms. As an AI language model, I am not qualified to provide medical diagnoses or treatment advice. However, based on your symptoms, there are a few possible causes:

- Acid reflux: The heartburn and nausea could be symptoms...
- Heart attack: Shoulder pain can sometimes be a symptom of...
- Gallbladder problems: Shoulder pain can also be a symptom...

In this example, the user disclosed critical health identifiers including age, weight, current medications for chronic conditions, and specific symptoms. This combination represents protected health information under both HIPAA and GDPR Article 9 regulations. According to the risk scoring framework we defined as part of the prompt construction process (see Section 4.2.1), for this conversation, the critical risk classification is correctly given due to the nature of the health disclosure and its potential for enabling discrimination or stigmatization in healthcare, employment, or insurance contexts.

Example 2: Business Contact Information with Geographic Indicators. In this conversation, the user requests legal document

assistance while inadvertently disclosing business contact information and geographic context. The LLM-based evaluation classified this conversation as high risk due to the combination of direct identifiers and location inference possibilities. According to the risk scoring framework, for this conversation, the LLM model assigned *Severity=2, Identifiability=2, and Exposure=4*. The privacy categories identified in this conversation are: CONTACT.EMAIL, LOC.ADDRESS. The associated LINDDUN threats are “*Identifiability*” and “*Disclosure*”, with the rationale that *business email addresses can directly identify individual operators and enable unwanted contact or harassment*.

Conversation Transcript 2

◀ **USER:**
 What information do you need to write a terms and conditions page for [COMPANY#D7B6C2F4] Fitness?

▶ **ASSISTANT:**
 To write a terms and conditions page for [COMPANY#D7B6C2F4] Fitness, we typically need... [services, payment terms, privacy policy...]

◀ **USER:**
 1) you know the business. 2) service is only a hormone optimization program designed to teach people to eliminate EDCs from their life. 3) use your own judgement 4) one time payment of [CURRENCY#F7A5D3B1] 999 inclusive of everything. 5) No refunds/cancellation 6) already have it on a separate page. 7) you do it 8) you do it 9) skip 10) [EMAIL#A7C6C4D1]

Now write the T&Cs for me

▶ **ASSISTANT:**
 Here’s a draft of the Terms and Conditions page for [COMPANY#D7B6C2F4] Fitness... Email: [EMAIL#A7C6C4D1]

The user disclosed a direct business email address that serves as a personal identifier, particularly relevant for small business operations where the email directly traces to an individual operator. The geographic context inference from currency denomination enhances location privacy risk by narrowing the user’s jurisdiction to his/her territories. The high exposure rating is justified as this contact information becomes publicly accessible through conversation logs, creating potential vectors for unwanted contact or business-targeted harassment under GDPR Article 4 personal data definitions.

F Validation Test Cases: Public vs. Private Information

It is a key concern in automated privacy risk assessment to ensure that the evaluation framework can correctly distinguish between references to publicly-available information and genuine disclosures of private data. For instance, a conversation discussing Donald Trump’s political positions should not be flagged as a privacy risk, while a conversation focused on someone’s personal health condition should be appropriately identified. This distinction requires contextual reasoning beyond simple pattern matching: the framework must understand subject types (public figure vs. private individuals), information sensitivity categories, and disclosure contexts.

Our evaluation leverages LLM capability to implement the above. To validate its capability, we designed a synthetic test suite of ten carefully constructed conversation scenarios (tc1 to tc10). The goal is to verify both aspects: (1) the framework correctly identifies genuine personal disclosures (true positives), and (2) it does not flag references to public information (true negatives).

We created ten synthetic conversations inspired upon and to simulate realistic AI chatbot interactions¹¹, organized into three categories:

I) Public Information Tests (True Negatives): These conversations reference publicly-known entities or general knowledge and are used to verify that the framework does not generate false positives. Test cases refer to:

- tc1 Political figures: a discussion of Trump, Obama, and Clinton’s campaign strategies
- tc2 Public companies: A technical comparison of Microsoft Azure, AWS, and Google Cloud services
- tc3 Tourist locations: Questions about visiting Paris, Tokyo, and Barcelona
- tc4 Family terminology: General questions about family relationship terms (uncle, aunt, cousin)
- tc5 Fictional character: Discussion of Walter White’s disease in the TV series Breaking Bad

II) Personal Information Tests (True Positives): These conversations involve genuine personal disclosures that should be detected and escalated appropriately. Test cases involve:

- tc6 Health disclosure: A discussion of medical conditions, with reference to medications, symptoms, and family health history
- tc7 Financial disclosure: Sharing of bank account credentials and financial details
- tc8 Contact disclosure: Email address, phone number, city, and LinkedIn profile

III) Contextual Complexity Tests: These scenarios test nuanced contextual reasoning capabilities. Test cases involve:

- tc9 Mixed context: Public employer (Stanford University) information, in combination with a disclosure of personal E-mail address
- tc10 Exposure dimension: A hypothetical question about SSN-format vs. an actual SSN disclosure

Each conversation was processed through our complete analysis pipeline, including pre-annotation for privacy preservation, LLM-based semantic analysis, and multi-dimensional risk scoring.

Results Overview. Table 9 summarizes the complete test results across all ten scenarios.

The framework achieved 100% specificity on public information: the framework did not raise any personal disclosure issues for conversations discussing Trump, Obama, Clinton (tc1), Microsoft, AWS, or Google Cloud (tc2), tourist locations (tc3), family relationship terminology (tc4), or fictional characters (tc5). This

¹¹Complete test materials are available in the supporting materials repository: <https://github.com/Majid-Mollaefar/chatbot-confessions-artifact>, including all synthetic conversation transcripts, detailed analysis results, and source code for full reproducibility.

Table 9: Validation test results demonstrating public vs. private distinction.

Test ID	Test Scenario	Category	#Findings	Risk	PII Codes Detected	Expected
<i>Public Information (True Negatives)</i>						
TC1	Political figures	Public	0	None	—	Yes
TC2	Cloud services	Public	0	None	—	Yes
TC3	Tourist locations	Public	2 [†]	High	LOC.ADDRESS [†]	Yes
TC4	Family terminology	Public	0	None	—	Yes
TC5	Fictional character	Public	0	None	—	Yes
<i>Personal Information (True Positives)</i>						
TC6	Health conditions	Personal	4	Critical	HEALTH.CONDITION, HEALTH.MEDICATION, PROFILE.DEMOGRAPHIC, PROFILE.FAMILY	Yes
TC7	Bank account	Personal	1	Critical	FIN.BANK_ACCOUNT	Yes
TC8	Contact info	Personal	1	High	PROFILE.NAME, CONTACT.EMAIL, CONTACT.PHONE, LOC.ADDRESS	Yes
<i>Contextual Complexity</i>						
TC9	Stanford + E-mail	Mixed	3	High	CONTACT.EMAIL, LOC.ADDRESS, PROFILE.PROFESSIONAL	Yes
TC10	Hypothetical vs. actual SSN	Exposure	2	Critical	ID.GOV'T, ID.GOV'T + PROFILE.DEMOGRAPHIC	Yes

[†]This test also included a home address mention, which was detected (2 High findings for contextual location indicators).

demonstrates that the framework does not indiscriminately flag all mentions of named entities but rather performs contextual reasoning about subject types and disclosure contexts.

For the personal information tests (true positives), the framework successfully detected all genuine privacy-sensitive disclosures across health, financial, and contact information categories. The health disclosure test (TC6) generated multiple findings spanning health conditions, medications, demographics, and family history, demonstrating the framework’s ability to identify interconnected sensitive information within a single conversation context. Financial credentials (TC7) and contact information (TC8) were similarly detected, confirming the framework’s sensitivity to direct personal identifiers.

The contextual complexity tests validated the framework’s nuanced reasoning capabilities. The mixed context scenario (TC9) correctly distinguished between public institutional affiliation (Stanford University) and personal contact details, flagging only the latter. The exposure dimension test (TC10) demonstrated the framework’s ability to differentiate between hypothetical discussions of sensitive topics versus actual disclosure of sensitive values, a critical distinction for practical privacy risk assessment.

Illustrative Examples. We present three representative examples with their complete analysis results.

Example 1: Public Figure Discussion (Zero Findings).

User conversation excerpt:

“I’m writing a political science essay about presidential campaign strategies. Can you help me analyze Donald Trump’s 2016 social media strategy? [...] Could you also explain how this compared to Barack Obama’s 2008 social media strategy?”

The conversation references public figures (Trump, Obama, Clinton) in a factual, academic context discussing publicly documented campaign strategies. The evaluation framework correctly classifies these as subject type *public figure* and highlights zero personal disclosure issues, demonstrating that it correctly distinguishes between public figures in factual discussions versus private disclosures about individuals.

Example 2: Health Information Disclosure.

User conversation excerpt:

“I need some medical advice. I’m a 45-year-old woman and I’ve been experiencing severe chest pain and shortness of breath for the past two days. I have a history of high blood pressure and my mother died of a heart attack at age 50. [...] I’m on Lisinopril 10mg daily for my blood pressure and I take metformin for my type 2 diabetes.”

Upon processing this conversation, the framework highlights multiple personal disclosures including HEALTH.CONDITION, HEALTH.MEDICATION, PROFILE.DEMOGRAPHIC, and PROFILE.FAMILY. It correctly applied automatic escalation for health-related PII and recognized that demographic information disclosed in a health context (age, gender) inherits elevated sensitivity due to the highly protected nature of medical health data.

Example 3: Hypothetical vs. Actual Disclosure

User conversation excerpt (two turns):

Turn 1: “I’m building a form validation system and need to understand different ID formats. Can you explain the format of US Social Security Numbers?”

Turn 2: “Thanks! [...] I need to test the form with real data. My SSN is 428-76-3921, and my date of birth is March 15, 1978. Can you verify this would pass validation?”

The framework correctly distinguishes between Turn 1 (discussing SSN format without revealing actual values) and Turn 2 (disclosing actual SSN and DOB, which lead to flagging both ID.GOV'T and PROFILE.DEMOGRAPHIC). This test case demonstrates the framework’s ability to distinguish *discussing* privacy-sensitive topics, and actually *disclosing* personal and sensitive attributes.