A Unified Framework for Prompt Privacy is Elusive and Misleading

Prakhar Ganesh[1*], **Yash More**[1*], **Marco Romanelli**[2], **Ferdinando Fioretto**[3], **Golnoosh Farnadi**[1]

[1]McGill University & Mila
[2]Hofstra University
[3]University of Virginia
prakhar.ganesh@mila.quebec, yash.more@alumni.ashoka.edu.in,
marco.romanelli@hofstra.edu, fioretto@virginia.edu, farnadig@mila.quebec

Abstract

As large language models (LLMs) become increasingly integrated into daily life, privacy concerns are on the rise. Driven by the appeal of convenient, universal solutions, current practices reflect a drift toward a one-size-fits-all approach to privacy. Unfortunately, we argue that a unified framework for prompt privacy is elusive and can instead mislead, creating even greater risk due to a false sense of safety. We identify five desirable properties of an effective privacy framework for prompts, namely protection guarantees, performance, efficiency, domain adaptability, and user accessibility, and highlight that existing frameworks, such as sanitization, differential privacy, cryptography, and contextual integrity, only satisfy a subset of these properties. Beyond individual frameworks, we find underlying tensions between these properties that preclude the development of a unified framework. We recommend two critical paths forward: emphasizing context-specific and application-specific evaluation of privacy frameworks, and fostering user awareness and privacy literacy.

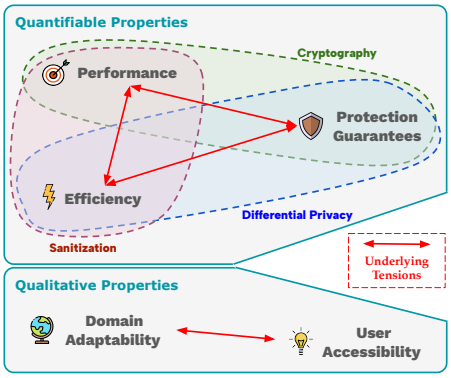

Figure 1: Five desirable properties of a prompt privacy framework (§ 2), coverage of various existing frameworks in the literature (§ 4, 5, 6), and several non-trivial underlying tensions between these properties (§ 8).

1 Introduction

Large language models (LLMs) have become increasingly embedded in our daily workflows, ranging from coding assistants and legal advisors to medical support tools and more (Zhao et al. 2025; Minaee et al. 2025; Chang et al. 2024). With their growing capabilities, LLMs have fundamentally changed how users interact with third-party commercial systems (Mireshghallah et al. 2024a; Touvron et al. 2023). This shift has led to a significant increase in the amount of sensitive information users share, whether due to the personalized and open-ended nature of conversations, better interactivity, or simply a general lack of AI literacy, particularly regarding how companies might collect and use their personal data (More, Ganesh, and Farnadi 2024; Carlini et al. 2021; Zhang et al. 2024b; Mireshghallah et al. 2024a).

Unsurprisingly, concerns around privacy in LLMs are on the rise (Du et al. 2025), both in academic research and in the design of tools for end users. Several solutions have been adopted from the existing privacy literature, including techniques like anonymization, differential privacy, and cryptography (Zhang et al. 2025; Hong et al. 2024; Shao et al. 2025; Du et al. 2023; Chen et al. 2023), while novel frameworks, such as contextual integrity (Mireshghallah et al. 2024b; Nissenbaum 2009), have also been proposed. The goal is often a generalizable, user-facing, plug-and-play framework that can be seamlessly integrated into any chatbot system.

Drift towards a Unified Framework: This evolution in prompt privacy reflects a broader trend toward unified solutions. Promises like *"Secure your AI. Everywhere it matters"* (pro 2025) or *"Privacy firewall for ChatGPT prompts"* (aimeetsprivacy 2025), without any explicit context, position these solutions as universal safeguards against all privacy risks. The diverse range of domains in which users engage with LLMs, ranging from everyday productivity tasks to financial, medical, or legal contexts, further compounds this trend. Consequently, users seek a universal solution to their privacy concerns. However, a *one-size-fits-all* approach to prompt privacy is deeply flawed.

Underlying Tensions in Prompt Privacy: We identify five key properties for prompt privacy (see Figure 1): three quantifiable (protection guarantees (Tong et al. 2025; Xin et al. 2025), performance (Sun et al. 2024; Chen et al. 2023), and efficiency (Edemacu and Wu 2025; Gim, Li, and Zhong 2024)) and two qualitative (domain adaptability (Mireshghallah et al. 2024b; Brown et al. 2022) and user accessibility (Zhang et al. 2025)). We find existing

[*]These authors contributed equally.

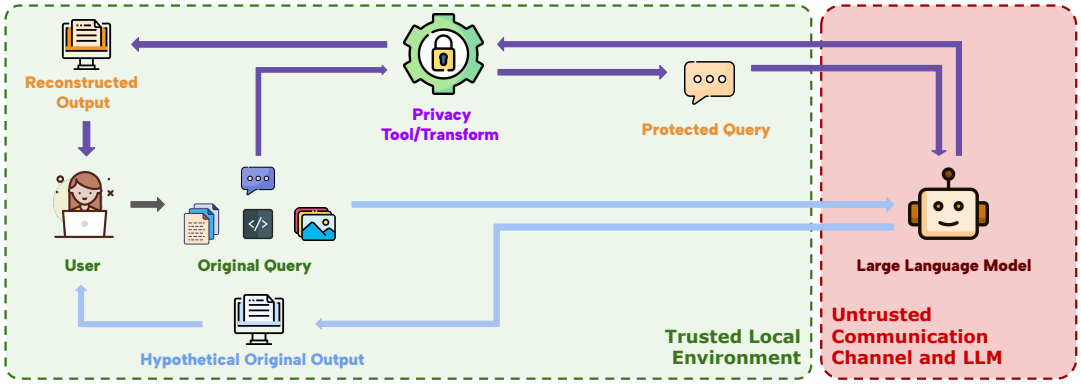

Figure 2: Prompt Privacy Pipeline. The user interacts with a privacy tool/transform before their query is sent to the LLM. All privacy interventions occur within a local trusted environment, ensuring that only the protected query is transmitted. We assume both the LLM and the communication channel can have malicious actors. The final reconstructed output is compared against the hypothetical output of the original query to make sure the LLM utility is preserved.

frameworks typically address only two of the three quantifiable properties, while also dealing with a trade-off between adaptability and accessibility. Digging deeper, we uncover two underlying tensions: first, the gap between natural language semantics and representation space, which complicates balancing protection, performance, and efficiency; and second, the wide range of LLM use cases, which makes adaptable frameworks capable of handling various privacy requirements inherently less accessible to everyday users.

We structure our argument as follows:

- We first define the scope of our discussion, and introduce five desirable properties of a privacy framework for prompts: protection guarantees, performance, efficiency, domain adaptability, and user accessibility (§ 2).
- We then examine four popular categories of privacy frameworks for prompts, i.e., sanitization (§ 4), differential privacy (§ 5), cryptography (§ 6), and contextual integrity (§ 7), highlighting the limitations of each approach.
- Moving beyond individual frameworks, we identify several underlying tensions between the desired properties (§ 8). These tensions represent non-trivial challenges that will likely persist even in future developments.
- We conclude with two key recommendations: context-specific evaluations to document the strengths and limitations of various frameworks across contexts, and better privacy literacy and awareness for the end users to help them choose and adopt the appropriate tool (§ 9).

2 Anatomy of a Prompt Privacy Framework

We start by defining a prompt privacy pipeline to establish the scope of our discussion. We then identify five desirable properties of a good privacy framework.

2.1 Prompt Privacy Pipeline

To establish a consistent terminology and mark the scope of our discussion, we present a prompt privacy pipeline in Figure 2. The pipeline contains the following components,

- **User:** The individual or entity interacting with the LLM.
- **Original Query:** The original input the user intends to send to the LLM. This includes both the language prompt and any auxiliary content such as documents, code, etc.
- **Privacy Tool/Transform:** The privacy tool responsible for protecting the original query and removing sensitive information, producing a protected query. In some cases, it also post-processes the LLM's response to reconstruct or recontextualize it before returning it to the user.
- **Protected Query:** The sanitized query generated by the privacy tool/transform.
- **LLM:** The target LLM to process the query.
- **Protected Output:** The response generated by the LLM when given the protected query as input.
- **Reconstructed Output:** The final response presented to the user, which may be the protected output or a refined version of it, depending on the privacy tool/transform.
- **Hypothetical Original Output:** The response generated by the LLM when given the original query as input.
- **Trusted Local Environment:** The environment that the user trusts. This need not be just the user's device; for example, if a trusted party, say the bank, intermediates access to an LLM, the trusted environment includes both the user's device and infrastructure under the bank's control.
- **Untrusted Communication Channel and LLM:** The communication channel through which the query is transmitted and the LLM itself, both assumed not to be trusted.

2.2 Desirable Properties of a Privacy Framework

We identify key properties that collectively determine the effectiveness and practical viability of a prompt privacy framework. These are not intended to be an exhaustive list, but rather a set of properties desirable for a *unified framework*.

Quantifiable Properties We start by defining three quantifiable properties fundamental to any privacy framework.

Protection Guarantees. Central to a privacy framework is its ability to protect against information leakage, i.e., limiting the information that an adversary can extract through the protected query. Depending on the context, the threat model, and the information available to the attacker, different strengths and nature of privacy protection might be desirable (Tong et al. 2025; Papadopoulou et al. 2022; Xin et al. 2025; Hong et al. 2024; Li et al. 2025b; Lu et al. 2023).

For instance, protection against honest-but-curious service providers may require different mechanisms than protection against malicious attackers with access to auxiliary information. Similarly, scenarios where attackers possess substantial background knowledge about users necessitate stronger privacy guarantees. Even the lifespan of stored data can influence privacy risks, as information that appears secure today may become vulnerable with more powerful adversaries in the future (Gomes, Sant'Ana, and Rodrigues 2025; Xin et al. 2025). Clearly, a framework's ability to provide protection guarantees is a critical property.

Performance. While protection guarantees are important, a good privacy framework must also maintain the performance of the LLM. Several different ways to quantify performance have been proposed in the literature, such as the semantic equivalence of the reconstructed output compared to the original output, the usability of the reconstructed output, or the preservation of factual accuracy and consistency in the reconstructed output, among others (Sun et al. 2024; Chen et al. 2023; Hong et al. 2024). More broadly, a good privacy framework should maintain output fidelity.

Efficiency. Efficiency relates to the computational aspects and latency of the privacy tool. A good privacy framework should minimize additional computational cost, avoid introducing significant latency, and scale effectively (Edemacu and Wu 2025; Gim, Li, and Zhong 2024). By keeping the processing lightweight and responsive, one can ensure that privacy protections do not hinder the use of the LLM system. Together, performance and efficiency focus on preserving the utility of the target LLM, thus maintaining the practical viability of the privacy framework.

Qualitative Properties Not all desirable properties of a privacy tool are easy to quantify. Here, we describe two qualitative properties of a good privacy framework.

Domain Adaptability. Real-world scenarios demand privacy frameworks that can flexibly adapt to diverse domains, even more so with the ever-growing list of LLM use cases (Mireshghallah et al. 2024b). This adaptability can take many forms, such as dynamically adjusting to various expectations of protection and types of adversaries, or being able to handle specialized language or formats across domains (Brown et al. 2022; Edemacu and Wu 2025).

For example, health-related queries or financial information often require stronger privacy protection than general knowledge questions or a casual conversation. In fact, some scenarios, like sharing proprietary information with an LLM, are more likely to involve a highly motivated adversary, thus requiring stricter privacy guarantees. Thus, the ability to adapt across various domains and LLM use cases is essential for an effective unified privacy framework.

User Accessibility. The successful adoption of any privacy framework depends on its accessibility to end users. This includes both the transparency of the framework as well as the cognitive burden of incorporating it into everyday workflows. When users can understand how their queries are transformed, they can verify that meaning has been preserved or debug cases where outputs are suboptimal, which can empower them to maintain appropriate protection. Notably, most widely adopted privacy tools tend to employ *sanitization* (pro 2025; aimeetsprivacy 2025; Chong et al. 2024; Zhang et al. 2025), a privacy framework that prioritizes accessibility. The appeal of a low barrier of adoption for the end user cannot be ignored.

Note that the two qualitative properties operate at a higher level than the quantifiable properties. Domain adaptability may be seen as a framework's ability to provide appropriate protection guarantees while maintaining performance and efficiency, across diverse domains. Similarly, user accessibility can be seen as the extent to which a framework makes the relationship between performance, efficiency, and privacy understandable and controllable for the end user.

3 User Profiles as Motivating Examples

Avery's Marathon Preparation. Avery wants to use an LLM chatbot to design a personalized fitness journey to help them prepare for a marathon. They aim to develop dietary plans, identify exercises, and create a strategy to be able to run the marathon in six months. As a meticulous researcher, Avery is not concerned about the accuracy of the LLM's suggestions, since they plan to verify the details independently. Their primary goal is to generate a quick initial draft of the plan that can be refined later. However, Avery is concerned about targeted ads if their chat history gets leaked accidentally or is sold intentionally by the third-party LLM.

Sasha's Work Companion. Sasha's company has rolled out a work companion using a third-party LLM API to assist employees with routine tasks such as summarizing reports, creating presentations, and more. Sasha plans to use this tool when working with confidential documents, so strong privacy protections are critical. A tool that requires constant supervision would be counterproductive, so Sasha will only use the work companion if it is highly accurate, even if it is costly. Sasha is not responsible for the compute costs, as the tool runs on the company's centralized infrastructure, while of course utilizing a third-party LLM API underneath. Note that the trusted local environment here includes the company infrastructure, but not the third-party LLM.

Tao's Writing Assistant. Tao is a non-native English speaker who wants to use an LLM-based writing assistant to help with personal emails and other forms of communication. Since these interactions are mostly casual, Tao is not overly concerned about privacy risks. However, they would like to ensure that no unnecessary identifying information is shared with the LLM. For Tao, accuracy and speed, allowing seamless integration of the tool into their daily life, are more important than strict privacy guarantees.

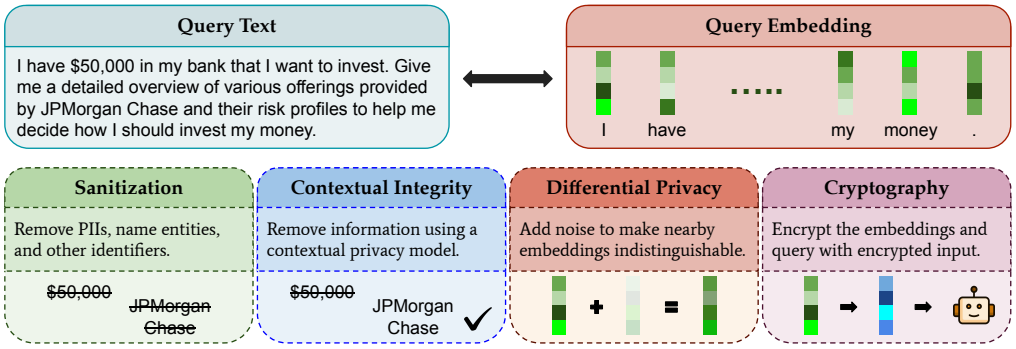

Figure 3: Four common frameworks for prompt privacy: sanitization, differential privacy, cryptography, and contextual integrity.

4 Sanitization: A Misplaced Sense of Safety

Sanitization refers to techniques that remove, mask, or generalize parts of the prompt in an effort to reduce exposure of sensitive information, as illustrated in Figure 3. These methods work at the surface level, making them efficient without degrading performance, but they lack rigorous protection guarantees and can create a false sense of safety. Sanitization often relies on lexical patterns and is thus highly accessible for end users. However, it can easily over- or under-sanitize since what counts as private varies widely across domains.

4.1 Lexical Sanitization misses Semantic Context

Lexical sanitization (Hedegaard, Houen, and Simonsen 2009; Papadopoulou et al. 2022; Albanese, Ciolek, and D'Ippolito 2023; Sánchez and Batet 2015) refers to techniques that focus solely on sanitizing specific words or phrases from a prompt, without considering the semantic meaning or broader context of the sentence. This category encompasses common approaches such as PII removal and named entity recognition (NER)-based methods. However, these techniques are frequently criticized for overlooking the complex linguistic and relational cues present in natural language (Guo, Shang, and Clavel 2024; Sinha et al. 2022). Moreover, since lexical sanitization methods rely on predefined lists of words or categories, they often fail to generalize across domains where sensitive information can take on many different forms. As a result, what qualifies as sensitive or identifying in one setting may go undetected in another, limiting their adaptability across a wide range of domains (Mireshghallah et al. 2024b).

Several works have shown that the syntactic properties of prompts, like paraphrasing, changes in formality, concreteness, grammatical mood, etc., can significantly influence the outputs of LLMs (Leidinger, van Rooij, and Shutova 2023; Rawte et al. 2023; Chataigner et al. 2025). In fact, the semantic interdependence of words remains hard to model even in information-theoretic frameworks, where it is necessary to have a finite set of valid alternatives to an invalid query (Glukhov et al. 2025). This further challenges the effectiveness of lexical sanitization, since these tools may fail to capture sensitive content if they miss key elements of the user's language. This disconnect between lexical cues and true context can instill a false sense of confidence in sanitization, leaving users exposed to privacy risks.

4.2 Lack of Guarantees under Sanitization

Recent works have revealed that the effectiveness of sanitization in removing sensitive information is often overestimated, leading to vulnerabilities that can be exploited by adversaries. Black-box attacks on LLMs have demonstrated that, even after sanitization or de-identification, adversaries can reconstruct or infer sensitive information from model outputs by exploiting residual semantic cues and contextual patterns (Carpentier et al. 2025; Hong et al. 2024; Li, Tan, and Liu 2025; Tong et al. 2025; Xin et al. 2025).

This vulnerability can extend beyond simple extraction to more sophisticated forms of reconstruction and inference, where attackers use contextual hints or auxiliary information to piece together private content. Such findings highlight that surface-level masking is rarely sufficient for robust privacy protection, especially in real-world scenarios with motivated adversaries (Li et al. 2025b). Hence, while sanitization can create a sense of privacy by removing certain words from the sentence, they have an even greater risk of misleading users (Edemacu and Wu 2025; Tong et al. 2025; Papadopoulou et al. 2022).

Sanitization frameworks offer a surface-level protection against privacy risks. While useful in some situations, sanitization cannot protect against motivated adversaries and fails to capture the dynamic, context-sensitive nature of real-world privacy across domains.
Sanitization risks exposing **Avery** to targeted ads and leaking company data for **Sasha**, making it unsuitable for them both. For **Tao**, however, it may be an appropriate tool, providing good performance and efficiency, along with sufficient surface-level privacy.

5 Differential Privacy: Promises and Pitfalls

Differential Privacy (DP) is a foundational framework of privacy in data-driven systems and has long been regarded as the gold standard in many traditional machine learning (ML) settings (Dwork and Roth 2014). At its core, DP provides a powerful guarantee: the output of a system should remain nearly indistinguishable regardless of the presence

or absence of a single individual in the data, achieved by introducing random noise. Thus, an adversary with access to the output cannot confidently predict the membership of an individual in the input data.

While DP is most commonly used to protect training datasets in ML, it can also be extended to prompt privacy. In this setting, DP is applied to the input prompt by injecting random noise through perturbations, ensuring that the perturbed prompt could have plausibly come from many different input prompts (Wu et al.), as illustrated in Figure 3. Thus, an adversary with access to only the perturbed prompt cannot confidently state which input prompt was used, thereby protecting information present in the original prompt.

5.1 Relative Bounds, Not Absolute Walls

DP provides relative bounds on information leakage rather than preventing the absolute flow of information. In essence, DP ensures that an adversary cannot reliably distinguish between neighboring prompts. However, if certain information is common across all neighbouring prompts, the adversary can still infer it. For example, suppose a user asks an LLM about a particular medication. DP may perturb the prompt so that the medication name is replaced with that of a similar but different drug. While this prevents the adversary from confidently recovering the original medication, even knowledge of the perturbed alternative still reveals significant information about the user's potential medical condition.

This can be partially addressed, somewhat artificially, by adjusting the strictness of DP through modifications to the definition of neighbouring. Expanding the notion of neighbouring increases protection, since indistinguishability must hold across a broader set of prompts. However, this also harms model performance, as essential information needed for accurate responses is suppressed. This tension gives rise to the well-known privacy-utility trade-off at the core of DP. Additionally, in the context of prompts, even defining what counts as neighbouring prompts introduces two fundamental challenges, which we explore in the following subsections.

5.2 The Structural Limits of DP for Prompts

DP relies on having a clear definition of a "secret" or sensitive record. For language data, and especially for prompts, it is difficult to establish consistent boundaries for what should be protected. Language does not have a universal unit for sensitive information. A "record" for DP can refer to a word, a sentence, a prompt, or even all the data from a single user.

On the one hand, if privacy is defined at the level of a prompt or an entire conversation, then the resulting guarantees are overly coarse and impractical, since the mechanism would need to inject prohibitive amounts of noise to provide formal protection, severely degrading utility. However, on the other hand, if privacy is defined at the word or sentence level, removing these units rarely hides private information, and does not prevent an adversary from inferring sensitive context from the rest of the prompt or related patterns (Brown et al. 2022; Mireshghallah et al. 2024a).

An interesting artifact of this can be seen in the existing literature that uses DP for prompts, where many prompt privacy techniques are designed for context-specific objectives, such as author obfuscation or single-sentence protection, rather than attempting to capture the broader, ever-evolving risks in real-world scenarios (Li et al. 2025b; Utpala, Hooker, and Chen 2023).

5.3 The Difficulty of Defining Prompt Similarity

Another challenge of defining neighbouring prompts in DP lies in choosing a meaningful similarity metric. As similarity in the semantic space is not directly quantifiable, it is often mapped into a different space, such as the embeddings. However, this introduces a gap between the measured similarity and the actual semantic similarity. As a result, these metrics frequently miss linguistic cues and stylistic markers through which sensitive information can leak. Consequently, even prompts that appear safe may contain identifying details, undermining the intended privacy guarantees (Du et al. 2023; Mattern, Weggenmann, and Kerschbaum 2022).

These difficulties are further compounded by the evolving nature of language and privacy norms. Language and the notion of what constitutes private or sensitive information are both fluid and continually evolving. Changes in linguistic usage, shifting social norms, and cultural movements can rapidly redefine how secrets are discussed or even whether certain information is seen as secret at all. On digital platforms, users frequently adopt coded language, euphemisms, or *algospeak* to bypass automated moderation, with new terms emerging as soon as old ones are flagged by algorithms. Any privacy approach, like DP, that relies on a static understanding of sensitive content will inevitably lag behind these changes, and thus fail to robustly protect user privacy in real time (Brown et al. 2022; Edemacu and Wu 2025).

Differential privacy offers strong guarantees, and can defend against malicious adversaries. However, its limitations are not just technical obstacles, but reflect a deeper misalignment between the rigid requirements of differential privacy and the fundamentally dynamic reality of language, which can render these guarantees obsolete in many real-world scenarios.

Performance drop under differential privacy would be unacceptable for both **Sasha** and **Tao**. While **Avery** could benefit from the protection provided by differential privacy, they should have a proper understanding of the privacy-utility trade-off to take advantage of DP.

6 Cryptography: A Costly Alternative

Cryptography in the context of privacy refers to a wide range of techniques that rely on secure computation, including encryption (Acar et al. 2018), secret sharing (Zhao et al. 2019), zero-knowledge proofs (Fiege, Fiat, and Shamir 1987), oblivious transfer (Yadav et al. 2022), etc. Two cryptographic techniques are common in prompt privacy literature: Homomorphic Encryption (HE) and Secure Multi-Party Computation (MPC). HE enables computation directly on encrypted text, as shown in Figure 3, and decrypts the output in the trusted environment (Zhang et al. 2024a;

Zimerman et al. 2025). MPC, on the other hand, distributes the computation across multiple independent parties, ensuring that no single party has access to the complete input (Xu et al. 2025). Cryptography can provide guarantees against information leakage, without altering the model input, thus preserving performance. However, it is computationally expensive and challenging to scale to LLMs.

6.1 The Barrier of Non-Linearity

HE holds great appeal because it promises the best of both worlds: privacy-preserving encryption and full utility of the model (Acar et al. 2018). In practice, however, it is highly constrained. HE is relatively efficient for linear algebra operations, such as the large matrix multiplications. The challenge arises with nonlinear operations like ReLU, softmax, and the attention mechanism itself, which can either make HE extremely slow or require polynomial approximations that quickly become unstable and expensive (Mittal 2024; Xu et al. 2025; Zhang et al. 2023). Thus, HE does not scale well for large neural networks, especially LLMs.

6.2 Specialized Infrastructure Requirements

A major barrier to the use of cryptographic techniques for prompt privacy in LLMs is the need for specialized infrastructure. MPC requires multiple non-colluding servers to jointly host and run the model, with heavy inter-server communication during inference. This is at direct odds with the way LLMs are deployed today, as centralized services optimized for single-node or tightly clustered GPU execution.

Similarly, HE also comes with significant infrastructure demands. Existing models trained on non encrypted text cannot be simply used "as is" on encrypted prompts. Instead, the model must be reimplemented within an HE-compatible framework, fundamentally changing the serving pipeline itself. In short, whether we use HE or MPC, the existing LLM infrastructure does not support these techniques out of the box, and would require substantial re-engineering.

6.3 Communication Overhead in Cryptography

Recent works have attempted to combine both HE and MPC, taking advantage of their strengths to create hybrid techniques (Xu et al. 2025; Lu et al. 2023; Pang et al. 2024). While this has provided significant improvements, these techniques do not escape the fundamental cost barrier. Even under optimistic assumptions of specialized infrastructure, hybrid techniques still introduce significant communication costs that accumulate quickly.

For instance, a common hybrid approach combines HE for linear layers with MPC for nonlinear layers (Xu et al. 2025). During inference, the client first encrypts the input and sends it to the server, where the linear computations of a single layer are done with HE. The encrypted output is then returned to the client, who decrypts it and then distributes it across multiple servers to handle the nonlinear computations using MPC. The outputs from MPC are combined to produce the final output. The entire pipeline is repeated for every subsequent layer of the model. This can significantly increase latency, communication data volume, or both. Consequently, although HE and MPC are powerful, they remain far too expensive for real-time, large-scale LLM services.

Cryptographic techniques for prompt privacy promise protection guarantees while maintaining model performance, however, fall short of true practical benefits due to their extremely high computational costs, limiting their use to toy models or shallow networks.

The high cost of cryptographic techniques makes it an unacceptable solution for **Avery** and **Tao**. As **Sasha's** work companion is deployed through the company infrastructure, they could benefit from the guarantees it provides against absolute information leakage. However, even for Sasha, a truly practical cryptographic system requires significant future developments.

7 Contextual Integrity: A Distant Ideal

Contextual Integrity (CI) (Nissenbaum 2004; Mireshghallah et al. 2024b) offers a principled way to analyze and adapt privacy norms as technology and society evolve. CI is the only framework in our discussion that was not adopted from traditional ML settings, and frames privacy not as a static property but as the appropriateness of information flow within a given social context, by examining who is sharing information, what is being shared, with whom, for what purpose, and under what transmission principles (Mireshghallah et al. 2024b; Shvartzshnaider and Duddu 2025; Nissenbaum 2004), an example shown in Figure 3.

While CI moves beyond binary public/private distinctions, recent critiques note that LLM research often "washes" CI in an attempt to operationalize it, adopting CI terminology while neglecting its core principles (Shvartzshnaider and Duddu 2025; Brown et al. 2022). Thus, despite its value as a theoretical lens to study privacy in prompts, translating it into practice remains a significant challenge. Several arguments in this section build on Shvartzshnaider and Duddu (Shvartzshnaider and Duddu 2025)'s position that CI is inadequately applied to LLMs.

7.1 Not Just a Special Case of Sanitization

Existing attempts to operationalize CI tend to encode fixed privacy norms or rules, drawing from legal codes, crowdsourced inputs, or author judgment (Shao et al. 2025; Ghalebikesabi et al. 2024; Cheng et al. 2024; Li et al. 2025a). This process simplifies a flexible, context-sensitive theory into a static set of templates or checklists. In fact, many approaches that operationalize CI do so through prompt sanitization or norm enforcement (Hartmann et al. 2024; Ngong et al. 2025; Shvartzshnaider and Duddu 2025).

As a result, these tools struggle to adapt to changing social contexts, evolving notions of sensitivity, and the diversity and dynamism inherent in real-world communication (Cofone 2018; Shvartzshnaider and Duddu 2025; Edemacu and Wu 2025). Consequently, they encounter practical barriers when essential and non-essential information is difficult to separate, or when attempts to sanitize prompts make them

unusable for the user's actual task (Shao et al. 2025). Rule-based methods are ill-suited for the open-ended, context-rich, and utility-driven scenarios that characterize prompt interactions with LLMs (Mireshghallah et al. 2024b).

Furthermore, even automated attempts to enforce CI, such as using LLMs or classifiers to judge appropriateness (Gu et al. 2025), add another layer of abstraction. These models often miss nuanced context, misclassify cases, or require significant user intervention. This tendency results in a cycle where the promise of robust privacy is not met in practice, and users are left with either reduced functionality or insufficient protection (Ngong et al. 2025).

7.2 The Difficulty of Defining Social Norms

CI is frequently reduced to regulatory compliance or minimization, losing the framework's distinctiveness (Shvartzshnaider and Duddu 2025), thus focusing on preventing the leakage of a fixed set of sensitive information or enforcing policy rules. However, true CI analysis should holistically assess the appropriateness of information flow in context, considering roles, values, and functions, not merely whether "private" information was shared (Shvartzshnaider and Duddu 2025; Yang et al. 2013; Barkhuus 2012). Laws and regulations, while influential, do not always reflect lived or moral norms in society, and privacy preferences collected from users may fail to capture the collective or ethical nature of privacy that CI intends to protect (Neel and Chang 2023).

The biggest culprit of this trend is the reduction of CI to focus almost exclusively on "sensitive information". This narrow focus leaves other privacy harms unaddressed, such as intrusion, profiling, or the gradual build-up of risk across a series of interactions. Privacy violations can occur over time or through patterns that single-turn evaluations do not capture, yet existing tools rarely address these longitudinal threats (Lukas et al. 2023).

Contextual integrity aims to study the dynamic nature of information flow in prompt privacy. However, translating this framework into a practical tool remains a challenge, without losing the very fluidity that makes it valuable. As Shvartzshnaider and Duddu (2025) argue, contextual integrity is a social theory, one that requires more than just purely algorithmic solutions.

The high barrier to entry makes contextual integrity an unlikely choice for **Tao**. For others, it could in principle be effective, for example, keeping **Avery** protected from targeted ads while still delivering a useful strategy, or preventing **Sasha** from sharing certain confidential documents. However, such context-specific tools should already exist and be accessible to the end users.

8 Underlying Tensions in Prompt Privacy

We have examined the shortcomings and challenges across various categories of existing privacy frameworks for prompts. However, even if these individual shortcomings were addressed in the future, we argue that there are underlying non-trivial tensions inherent to prompt privacy in the age of LLMs that would prevent the development of a unified framework. We discuss these tensions below.

8.1 Guarantees, Performance and Efficiency

Regardless of the underlying model, natural language processing (NLP) systems transform language into vector representations through embeddings. This transformation from natural language to representation creates an inherent disconnect that generates tension between three critical dimensions: protection guarantees, performance, and efficiency.

The core challenge stems from defining "similarity" across two spaces. Semantic similarity in language space may not always align with similarity in the representational space (§ 5.3), affecting both protection guarantees and LLM utility (i.e., performance and efficiency). Privacy frameworks that focus on the representations, such as differential privacy (§ 5) and cryptography (§ 6), operate in a mathematical domain where information leakage guarantees can be formally proven. Yet, to truly prevent sensitive information flow, these techniques need to 'over-protect' at the representational level, to ensure appropriate protection in the language space, at the cost of degraded performance (differential privacy, § 5) or reduced efficiency (cryptography, § 6). On the other hand, frameworks that operate in the language space, such as sanitization (§ 4), can maintain performance and efficiency, but lack any protection guarantees.

This is more than just a trivial limitation of existing privacy frameworks. Any future privacy framework must also deal with the underlying tension between choosing stronger guarantees in the representational space (at the cost of performance or efficiency) or better performance and efficiency in language space (at the cost of formal guarantees).

8.2 User Accessibility and Domain Adaptability

Beyond the three quantifiable properties, we also defined two qualitative properties of a good prompt privacy framework (§ 2.2). Through the study of existing frameworks, we find another underlying tension in prompt privacy between these two qualitative properties. The tension between user accessibility and domain adaptability emerges from the remarkable breadth of applications for which LLMs are now employed. This diversity of use cases creates vastly different privacy requirements across domains (§ 4.1, 7.2). For instance, medical consultations, legal document review, creative writing, and casual conversation each demand distinct privacy considerations based on different notions of sensitivity, regulatory requirements, and social norms.

Unfortunately, a framework capable of truly adapting to this variety of domains and covering all different requirements necessarily demands significant privacy literacy from users. Thus, the accessibility and ease of use of a privacy framework stand in fundamental opposition to the comprehensive adaptability required for LLM applications. This creates an unavoidable choice: frameworks can either be accessible and easy to use, or they can be highly adaptable across domains, but not both.

Accessible frameworks like sanitization (§ 4) require minimal cognitive overhead, and users can quickly apply these

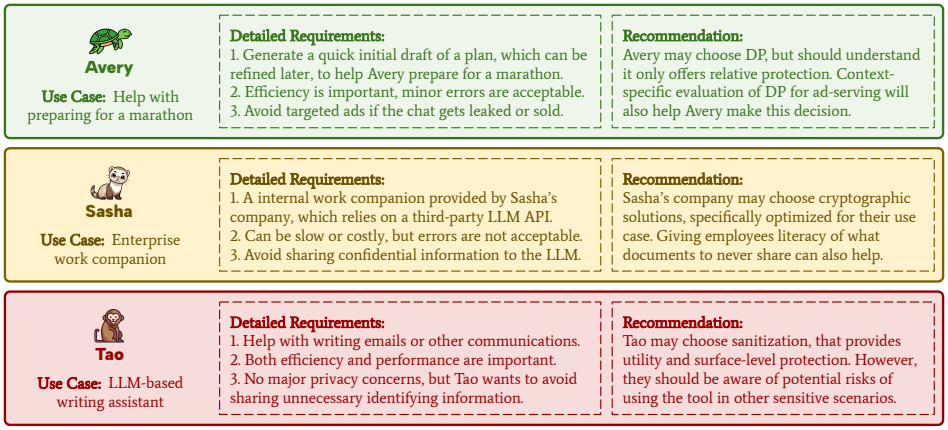

Figure 4: Motivating use cases as defined in § 3, along with recommendations based on our discussions in the paper.

approaches without extensive training. However, they struggle with domain adaptability due to changing specialized terminology across fields, evolving social norms, and varying definitions of what constitutes sensitive information in different contexts (§ 4.1). On the other end of the spectrum, adaptable frameworks like contextual integrity (§ 7) are specifically designed for cross-domain flexibility. However, they demand domain-specific expertise, a deep understanding of how privacy situations are constructed, and how sensitivity is defined in various contexts (§ 7.2). This complexity creates significant barriers to adoption by end users.

9 Where do we go from here?

We argued that a unified framework for privacy is elusive, and that trying to promote one as a universal solution will mislead users. However, we bring the reader's attention back to motivating examples in § 3. Throughout our discussion, we found that even though unified frameworks are out of reach, certain tools can benefit theseusers in concrete scenarios. We summarize these solutions in Figure 4. Two key points deserve emphasis and form the basis of our recommendations: first, effective context-specific solutions are possible; second, improving end users' privacy literacy is essential, both to help them select the appropriate solution and to understand its limitations.

Context-Specific Evaluations. We find that the first step toward building an effective privacy tool is recognizing the requirements of the application. Understanding what the user needs, what trade-offs they will accept, and which privacy risks are relevant helps guide the selection of an appropriate tool. Theories like contextual integrity (Nissenbaum 2004) can be particularly useful in this process. The *"Detailed Requirements"* in Figure 4 illustrates this exercise.

Once the requirements are identified, the next challenge is assessing how well different privacy frameworks perform in that setting. For example, based solely on the fundamental properties of DP, it is unclear whether it can provide sufficient protection against targeted ads that Avery might receive. Similarly, the boundary of when sanitization can be problematic in email writing for Tao remains poorly understood. While many works focus on improving privacy frameworks, evaluating these systems in specific contexts is rare.

We strongly recommend that the community move toward context-specific and application-specific evaluation of privacy frameworks, to gain a clearer understanding of their strengths and limitations for real-world use.

Fostering Privacy Literacy. In the absence of a universal solution and with the search for context-specific privacy frameworks, it becomes even more important to help users understand the trade-offs involved in using different tools. For instance, while no single tool is ideal for Sasha's company (cryptography being the closest, though still impractical), educating employees about which types of company information should never be shared with an LLM can help reduce information leakage while still providing them with an LLM-based work companion.

Many users remain unaware of how their inputs may be stored, aggregated, or used for training, which can create a false sense of safety and lead to uninformed sharing of information. Future systems should explore mechanisms to communicate these risks proactively. This could include in-context warnings, real-time feedback on risky disclosures, and educational efforts to improve literacy around privacy.

Empowering users with stronger literacy and accurate mental models of how their data is handled is essential to fostering safer and trustworthy interactions.

Prompt privacy in LLMs will remain a moving target, shaped by its expanding list of applications and user expectations. Rather than chasing a unified framework, we argue for context-aware solutions coupled with efforts to raise privacy literacy, which can together make meaningful progress toward safer and more trustworthy systems.

Acknowledgment

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, MITACS, FRQNT, and NSERC Discovery Grants program. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

References

2025. Prompt Security: AI Security Company — Manage GenAI Risks & Secure LLM Apps. https://www.prompt.security/. Accessed: 2025-09-14.

Acar, A.; Aksu, H.; Uluagac, A. S.; and Conti, M. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4): 1–35.

aimeetsprivacy. 2025. PrivacyGPT. Chrome Web Store extension. Version 1.3.2, last updated May 29, 2025; offered by aimeetsprivacy.

Albanese, F.; Ciolek, D.; and D'Ippolito, N. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785*.

Barkhuus, L. 2012. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 367–376.

Brown, H.; Lee, K.; Mireshghallah, F.; Shokri, R.; and Tramèr, F. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2280–2292.

Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, Ú.; Oprea, A.; and Raffel, C. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650. USENIX Association. ISBN 978-1-939133-24-3.

Carpentier, R.; Zhao, B. Z. H.; Asghar, H. J.; and Kaafar, D. 2025. Preempting Text Sanitization Utility in Resource-Constrained Privacy-Preserving LLM Interactions. arXiv:2411.11521.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Chataigner, C.; Ma, R.; Ganesh, P.; Taïk, A.; Creager, E.; and Farnadi, G. 2025. Say It Another Way: A Framework for User-Grounded Paraphrasing. arXiv:2505.03563.

Chen, Y.; Li, T.; Liu, H.; and Yu, Y. 2023. Hide and seek (has): A lightweight framework for prompt privacy protection. *arXiv preprint arXiv:2309.03057*.

Cheng, Z.; Wan, D.; Abueg, M.; Ghalebikesabi, S.; Yi, R.; Bagdasarian, E.; Balle, B.; Mellem, S.; and O'Banion, S. 2024. Ci-bench: Benchmarking contextual integrity of ai assistants on synthetic data. *arXiv preprint arXiv:2409.13903*.

Chong, C. J.; Hou, C.; Yao, Z.; and Talebi, S. M. S. 2024. Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. arXiv:2408.07004.

Cofone, I. N. 2018. Privacy Harms. *Hastings Law Journal*, 69(4): 1039–1070.

Du, H.; Liu, S.; Zheng, L.; Cao, Y.; Nakamura, A.; and Chen, L. 2025. Privacy in Fine-tuning Large Language Models: Attacks, Defenses, and Future Directions. arXiv:2412.16504.

Du, M.; Yue, X.; Chow, S. S. M.; Wang, T.; Huang, C.; and Sun, H. 2023. DP-Forward: Fine-tuning and Inference on Language Models with Differential Privacy in Forward Pass.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4): 211–407.

Edemacu, K.; and Wu, X. 2025. Privacy preserving prompt engineering: A survey. *ACM Computing Surveys*, 57(10): 1–36.

Fiege, U.; Fiat, A.; and Shamir, A. 1987. Zero knowledge proofs of identity. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, 210–217.

Ghalebikesabi, S.; Bagdasaryan, E.; Yi, R.; Yona, I.; Shumailov, I.; Pappu, A.; Shi, C.; Weidinger, L.; Stanforth, R.; Berrada, L.; et al. 2024. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*.

Gim, I.; Li, C.; and Zhong, L. 2024. Confidential prompting: Protecting user prompts from cloud llm providers. *arXiv preprint arXiv:2409.19134*.

Glukhov, D.; Han, Z.; Shumailov, I.; Papyan, V.; and Papernot, N. 2025. Breach By A Thousand Leaks: Unsafe Information Leakage in 'Safe' AI Responses. In *The Thirteenth International Conference on Learning Representations*.

Gomes, A. G.; Sant'Ana, R. C. G.; and Rodrigues, F. d. A. 2025. The mosaic effect on user identification: Implications for privacy. *Encontros Bibli*, 30: e102723.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; Wang, S.; Zhang, K.; Wang, Y.; Gao, W.; Ni, L.; and Guo, J. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Guo, Y.; Shang, G.; and Clavel, C. 2024. Benchmarking Linguistic Diversity of Large Language Models. arXiv:2412.10271.

Hartmann, F.; Tran, D.-H.; Kairouz, P.; Cărbune, V.; and Arcas, B. A. Y. 2024. Can LLMs get help from other LLMs without revealing private information? In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, 107–122.

Hedegaard, S.; Houen, S.; and Simonsen, J. G. 2009. Lair: a language for automated semantics-aware text sanitization based on frame semantics. In *2009 IEEE International Conference on Semantic Computing*, 47–52. IEEE.

Hong, J.; Wang, J. T.; Zhang, C.; Li, Z.; Li, B.; and Wang, Z. 2024. DP-OPT: Make Large Language Model Your Privacy-Preserving Prompt Engineer. arXiv:2312.03724.

Leidinger, A.; van Rooij, R.; and Shutova, E. 2023. The language of prompting: What linguistic properties make a prompt successful? arXiv:2311.01967.

Li, H.; Hu, W.; Jing, H.; Chen, Y.; Hu, Q.; Han, S.; Chu, T.; Hu, P.; and Song, Y. 2025a. Privaci-bench: Evaluating privacy with contextual integrity and legal compliance. *arXiv preprint arXiv:2502.17041*.

Li, M.; Fan, H.; Fu, S.; Ding, J.; and Feng, Y. 2025b. DP-GTR: Differentially Private Prompt Protection via Group Text Rewriting. arXiv:2503.04990.

Li, Y.; Tan, Z.; and Liu, Y. 2025. Privacy-Preserving Prompt Tuning for Large Language Model Services. arXiv:2305.06212.

Lu, W.-j.; Huang, Z.; Gu, Z.; Li, J.; Liu, J.; Hong, C.; Ren, K.; Wei, T.; and Chen, W. 2023. Bumblebee: Secure two-party inference framework for large transformers. *Cryptology ePrint Archive*.

Lukas, N.; Salem, A.; Sim, R.; Tople, S.; Wutschitz, L.; and Zanella-Béguelin, S. 2023. Analyzing Leakage of Personally Identifiable Information in Language Models. arXiv:2302.00539.

Mattern, J.; Weggenmann, B.; and Kerschbaum, F. 2022. The Limits of Word Level Differential Privacy. arXiv:2205.02130.

Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; and Gao, J. 2025. Large Language Models: A Survey. arXiv:2402.06196.

Mireshghallah, N.; Antoniak, M.; More, Y.; Choi, Y.; and Farnadi, G. 2024a. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. arXiv:2407.11438.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2024b. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. arXiv:2310.17884.

Mittal, R. 2024. Improving Inference Privacy for Large Language Models using Fully Homomorphic Encryption.

More, Y.; Ganesh, P.; and Farnadi, G. 2024. Towards More Realistic Extraction Attacks: An Adversarial Perspective. arXiv:2407.02596.

Neel, S.; and Chang, P. 2023. Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

Ngong, I.; Kadhe, S.; Wang, H.; Murugesan, K.; Weisz, J. D.; Dhurandhar, A.; and Ramamurthy, K. N. 2025. Protecting Users From Themselves: Safeguarding Contextual Privacy in Interactions with Conversational Agents. arXiv:2502.18509.

Nissenbaum, H. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79: 119.

Nissenbaum, H. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. USA: Stanford University Press. ISBN 0804752370.

Pang, Q.; Zhu, J.; Möllering, H.; Zheng, W.; and Schneider, T. 2024. Bolt: Privacy-preserving, accurate and efficient inference for transformers. In *2024 IEEE Symposium on Security and Privacy (SP)*, 4753–4771. IEEE.

Papadopoulou, A.; Yu, Y.; Lison, P.; and Øvrelid, L. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 217–229.

Rawte, V.; Priya, P.; Tonmoy, S. M. T. I.; Zaman, S. M. M.; Sheth, A.; and Das, A. 2023. Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness. arXiv:2309.11064.

Shao, Y.; Li, T.; Shi, W.; Liu, Y.; and Yang, D. 2025. PrivacyLens: Evaluating Privacy Norm Awareness of Language Models in Action. arXiv:2409.00138.

Shvartzshnaider, Y.; and Duddu, V. 2025. Position: Contextual Integrity is Inadequately Applied to Language Models. arXiv:2501.19173.

Sinha, K.; Gauthier, J.; Mueller, A.; Misra, K.; Fuentes, K.; Levy, R.; and Williams, A. 2022. Language model acceptability judgements are not always robust to context. arXiv:2212.08979.

Sun, X.; Liu, G.; He, Z.; Li, H.; and Li, X. 2024. DePrompt: Desensitization and Evaluation of Personal Identifiable Information in Large Language Model Prompts. *arXiv preprint arXiv:2408.08930*.

Sánchez, D.; and Batet, M. 2015. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1): 148–163.

Tong, M.; Chen, K.; Yuan, X.; Liu, J.; Zhang, W.; Yu, N.; and Zhang, J. 2025. On the Vulnerability of Text Sanitization. arXiv:2410.17052.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Utpala, S.; Hooker, S.; and Chen, P.-Y. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8442–8457. Singapore: Association for Computational Linguistics.

Wu, H.; Dai, W.; Li, W.; and Yan, Q. ???? Cape: Context-Aware Prompt Perturbation Mechanism with Differential Privacy. In *Forty-second International Conference on Machine Learning*.

Xin, R.; Mireshghallah, N.; Li, S. S.; Duan, M.; Kim, H.; Choi, Y.; Tsvetkov, Y.; Oh, S.; and Koh, P. W. 2025. A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage. arXiv:2504.21035.

Xu, T.; Lu, W.-j.; Yu, J.; Chen, Y.; Lin, C.; Wang, R.; and Li, M. 2025. Breaking the Layer Barrier: Remodeling Private Transformer Inference with Hybrid {CKKS} and {MPC}. In *34th USENIX Security Symposium (USENIX Security 25)*, 2653–2672.

Yadav, V. K.; Andola, N.; Verma, S.; and Venkatesan, S. 2022. A survey of oblivious transfer protocol. *ACM Computing Surveys (CSUR)*, 54(10s): 1–37.

Yang, X.; Wen, Z.; Qu, W.; Chen, Z.; Xiang, Z.; Chen, B.; and Yao, H. 2013. Memorization and Privacy Risks in Domain-Specific Large Language Models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Zhang, J.; Chen, K.; Feng, Z.; Lou, J.; and Song, M. 2024a. SecPE: Secure Prompt Ensembling for Private and Robust Large Language Models. In *ECAI*.

Zhang, L.; Li, C.; Hu, Q.; Lang, J.; Huang, S.; Hu, L.; Leng, J.; Chen, Q.; and Lv, C. 2023. Enhancing privacy in large language model with homomorphic encryption and sparse attention. *Applied Sciences*, 13(24): 13146.

Zhang, S.; Yi, X.; Xing, H.; Ye, L.; Hu, Y.; and Li, H. 2025. Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction. arXiv:2410.15044.

Zhang, Z.; Jia, M.; Lee, H.-P.; Yao, B.; Das, S.; Lerner, A.; Wang, D.; and Li, T. 2024b. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–26.

Zhao, C.; Zhao, S.; Zhao, M.; Chen, Z.; Gao, C.-Z.; Li, H.; and Tan, Y.-a. 2019. Secure multi-party computation: theory, practice and applications. *Information Sciences*, 476: 357–372.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2025. A Survey of Large Language Models. arXiv:2303.18223.

Zimerman, I.; Adir, A.; Aharoni, E.; Avitan, M.; Baruch, M.; Drucker, N.; Masalha, R.; Meiri, R.; and Soceanu, O. 2025. PowerSoftmax: Towards secure LLM Inference Over FHE. In *Annual FHE. org Conference on Fully Homomorphic Encryption.*