Balancing Transparency and Privacy in AI

Vidhi Bhatt[1], **Michael Shi**[1], **Fred Xu**[1], **Salman Rahman**[1]

[1]University of California, Los Angeles (UCLA)

Correspondence: vidhib@cs.ucla.edu, michaelshi@ucla.edu, fredxu@cs.ucla.edu, salman@cs.ucla.edu

Abstract

Large language models have become widely adopted and used both commercially and personally, which has led to heightened scrutiny around data privacy and transparency. Many AI systems process vast amounts of user data, but it is often unclear how this data is collected, stored, and used. Ensuring transparency in AI models is essential not just for accountability and trust, but also to prevent potential harm to users and others who may not even know that their privacy has been infringed upon. In our project, we will explore these issues by examining the challenges discussed in What Does it Mean for a Language Model to Preserve Privacy (Brown et al., 2022)? This paper highlights the difficulties of protecting user data while also making AI systems more understandable and open. Our goal is to develop a policy framework that helps strike a balance between these two priorities. We will consider different strategies, including legal frameworks and possible regulatory solutions, technical solutions like differential privacy, and best practices for model training and data usage. Our framework will aim to provide practical recommendations that address both ethical and legal concerns while ensuring that AI can be developed effectively yet responsibly. By tackling these challenges, we hope to contribute to the ongoing discussion on responsible AI development and offer insights for policymakers, researchers, and organizations working with AI technology.

1 Introduction and Motivation

Privacy has become a critical concern in the digital age, particularly as large-scale data collection and analysis have become central to various domains, including artificial intelligence (AI), finance, and healthcare. The rapid deployment of large language models (LLMs) in commercial and personal applications has created an urgent tension between transparency and privacy. While users increasingly demand to understand how AI systems process their data, these same transparency mechanisms can potentially reveal sensitive information.

The paper *What Does it Mean for a Language Model to Preserve Privacy* (Brown et al., 2022) highlights the fundamental challenges in conceptualizing, defining, and enforcing privacy in a meaningful and effective manner. Despite existing privacy frameworks such as differential privacy and legal regulations like the General Data Protection Regulation (GDPR), ensuring comprehensive privacy protection remains a complex task due to the contextual nature of language, the difficulty in identifying secrets, the evolving use of language, and fundamental differences between individual and group privacy.

Brown et al. (Brown et al., 2022) identified critical limitations in current approaches: traditional data sanitization is insufficient because it (a) requires tokens to be structured properly; (b) fails to account for contexts; (c) includes secrets when partial information is repeated; and (d) fails to adapt to the evolution of language usage. Moreover, the granularity of secrets presents challenges, as what constitutes private information varies across individuals and groups. This makes rigorous methods like differential privacy problematic to enforce, since group privacy guarantees often decrease exponentially with group size.

These limitations demand a new policy framework that balances transparency with robust privacy protections. In this report, we address these challenges by synthesizing insights from more recent research on context-aware privacy mechanisms, private pre-training approaches, contextual integrity in LLMs, and unlearning techniques. We propose domain-specific privacy policies that incorporate both technical solutions and normative contexts, aiming to provide actionable recommendations for AI developers, policymakers, and organizations deploying AI systems. Recent work has expanded on many of the key issues raised by Brown et al., particularly in the context of emerging AI technologies, decentralized computing, and cryptographic approaches to privacy.

2 Related Works

We examine recent research that addresses the privacy challenges identified by (Brown et al., 2022), focusing on four key areas: context-aware differential privacy, contextual integrity in LLMs, privacy considerations in pre-training, and unlearning mechanisms.

2.1 Context-Aware Differential Privacy

Dinh and Fioretto (2023) proposed the context of context-aware differential privacy (CADP). The authors correctly recognize the important role of context and points out that sensitive information may be described in multiple, often ambiguous, ways and context is often useful to infer their sensitivity. To reason about context more clearly, the authors define contexts as a subsequence of tokens such that under an invariant mapping ϕ that transforms sequences into other sequences with similar semantics, the probability of LM generating the next token as sensitive (private) changes within a pre-defined value of α. e.g., a sequence $\tilde{x}$ is called α-context of sensitive information x_i, preceeded by sequence $x = \{x_1, \cdots, x_i\}$, if $|P(x_i \mid \phi(\tilde{x})) - P(x_i \mid x)| \leq \alpha$.

The main contributions of this work, from the perspective of (Brown et al., 2022), are that (1) it clearly defined what context means for sensitive information, and (2) it proposed a way to enforce differential privacy with contexts, directly addressing challenge (b) from the introduction section. This definition enables the design of training algorithms for differentially private language models that account for contextual information.

Building on this work, (Benthall and Cummings, 2024) highlighted the lack of normative aspects in differential privacy approaches and bridged this gap with the concept of *Contextual Integrity (CI)*. This framework incorporates normative social expectations across specific domains like healthcare, financial systems, and education. Beyond the technical situation of the system (which includes populations involved, distribution of heterogeneous features, and applicable threat models), the framework adds normative spheres to provide additional context. By considering both normative and descriptive aspects of social environments and technology in one unified framework, this approach offers a more comprehensive view of privacy. In Section 3, we will propose potential policies based on the insights from both (Dinh and Fioretto, 2023) and (Benthall and Cummings, 2024).

2.2 Contextual Integrity and Privacy Reasoning in LLMs

Mireshghallah et al. (2023) extends the contextual understanding of privacy by examining how LLMs handle sensitive information at inference time through CONFAIDE, a multi-tiered benchmark grounded in contextual integrity theory. Their experiments expose significant weaknesses in models' privacy reasoning capabilities, showing even heavily aligned commercial models like GPT-4 reveal private information in contexts where humans wouldn't (39% of the time in real-world scenarios).

Shvartzshnaider and Duddu (2025) identify a concerning trend they term "CI-washing," where researchers claim to use contextual integrity frameworks while neglecting its core tenets, particularly the evaluation of ethical legitimacy through the CI heuristic. This leads to privacy evaluations that focus excessively on binary public/private classifications rather than appropriateness of information flows within specific social contexts, highlighting the need for more rigorous application of privacy theory to LLM development.

2.3 Privacy Considerations in Pre-training

Tramèr et al. (2024) examine privacy issues in model pre-training in their work *Position: Considerations for Differentially Private Learning with Large-Scale Public Pretraining*. They analyze both the performance of differentially private machine learning and privacy considerations of public datasets used for pre-training. Their research highlights that "data scraped from the Web may be sensitive itself," allowing a "privacy-preserving finetuned model to still memorize its pretraining data."

This presents a fundamental challenge: despite using privacy protections during fine-tuning, models can still expose sensitive information from their pre-training data that was scraped publicly. Even "public" data may contain sensitive information shared without the original party's consent or knowledge. Without universal consent guidelines, it becomes nearly impossible to verify privacy guarantees in pre-training datasets, which consequently complicates ensuring differential privacy in models built on this data.

The authors also identify a performance issue: when fine-tuning these models on downstream private data, results are often inconsistent due to misalignment between public and private datasets. This highlights the need for better private-data benchmarks to separate public and private data concerns, ensuring private models are trained and evaluated on private data from end to end while public-data models maintain their effectiveness in public settings.

2.4 Unlearning Mechanisms and Benchmarks

A critical challenge in privacy-preserving language models is the process of unlearning—removing specific data from a model after training. This directly addresses point (c) in the introduction: ensuring secrets are not retained when partial information is repeated. Thaker et al. (2024) in their paper *Position: LLM Unlearning Benchmarks are Weak Measures of Progress* critique existing methods of measuring unlearning in LLMs, arguing that current benchmarks fail to capture meaningful progress.

The authors identify two key deficiencies in current unlearning benchmarks. First, most evaluate unlearning by testing whether a model forgets specific data points rather than assessing broader systemic risks of memorization. Second, these benchmarks inadequately measure generalization of forgetting—whether a model not only forgets explicitly identified data but also avoids inferring sensitive patterns from related information. The paper recommends that new benchmarks should focus on the structural aspects of how LLMs retain and generalize sensitive information.

3 Discussion and Policies

3.1 Incorporating Context-aware Differential Privacy Practices

From (Dinh and Fioretto, 2023; Benthall and Cummings, 2024), we can draw two conclusions: (1) definition of differential privacy should consider what a context is on the algorithmic level, and (2) this context should also be extended to the non-algorithmic level, namely normative. On the policy-level, regulation bodies and especially corporations, when rolling out machine learning models to production, should consider themselves with the following questions when applying differential privacy: regarding (1), is the current model trained using differential privacy, and if so, does it incorporate contexts like the one defined in (Dinh and Fioretto, 2023)? Regarding (2), how are the hyperparameters of the DP determined? Based on what normative contexts, if any, were these hyperparameters determined? If the answer to the two questions are negative or unclear, it is necessary for the corporation to start an auditing process that addresses them.

3.2 Improving the future of private pre-training

As recommended in (Tramèr et al., 2024), there are four primary directions in which future research can improve private training in order to develop differentially private models that can be effective at the scale of large public models:

- Generating datasets from publicly scraped web data should require greater granularity to ensure consent and privacy when it comes to sensitive data. With additional metadata or tags on lower levels, researchers can have greater understanding and assurance in whether or not data is sensitive or not and whether or not it is being gathered with consent.

- The next step is from the models themselves: privacy-friendly pretrained models that gather data should use express consent-of-use to ensure differential privacy.

- Benchmarks should be made that mimic private datasets for downstream tasks that involve sensitive information to better evaluate these privacy-friendly models. A lack of such benchmarks makes private learning additionally difficult.

- As seen from the previous three directions, a great focus is put on the information flow and data. Further research should be more data-centric as opposed to model-centric to ensure that privacy is ensured throughout the flow of information from end to end.

It is clear that privacy needs to be ensured from end-to-end, which make pretraining privacy insurance and a greater focus on the data that flows through models vital. Improved data gathering techniques and private benchmarks will support frameworks such as those we discuss in this paper and beyond. We would further like to add that in addition to these directions, our insights related to contextually-aware differential privacy and contextual integrity inform that research in pre-training should also take contexts into account when it comes to assessing data.

3.3 Beyond Surface-Level Privacy: The Challenge of Contextual Integrity in LLMs

Building on the "CI-washing" phenomenon identified by Shvartzshnaider and Duddu (2025), where researchers claim to use Contextual Integrity frameworks while neglecting core principles, we recognize the need for policy interventions that address these methodological gaps. Particularly concerning is how evaluations of LLMs through a CI lens routinely fail to account for inherent variabilities—including prompt sensitivity, position bias, and varying responses to identical queries—potentially invalidating conclusions about LLMs' privacy capabilities.

For policymakers, these findings highlight the urgent need for more nuanced privacy frameworks that go beyond binary classifications of data. Based on the shortcomings in current implementations of CI theory, we propose the following policy recommendations:

- Develop standardized frameworks for CI-based privacy assessments that require explicit consideration of all four tenets, with particular emphasis on defining contextual norms
- Mandate that privacy evaluations of commercial LLMs include input from domain experts who understand contextual norms in specific social spheres (healthcare, education, finance)
- Create sector-specific privacy requirements for LLMs that reflect the unique contextual norms of different domains rather than applying generic privacy standards
- Establish minimum standards for experimental robustness in LLM privacy evaluations, requiring controls for position bias and prompt sensitivity
- Require explicit disclosure of which CI tenets are being implemented when systems claim to be evaluating privacy through contextual integrity

3.4 Increasing robustness of AI privacy regarding unlearning

The paper *Position: LLM Unlearning Benchmarks are Weak Measures of Progress* (Thaker et al., 2024) argues that simply deleting data from a model doesn't ensure real privacy, as models don't just memorize—they infer patterns and relationships that can still expose sensitive information. (Thaker et al., 2024) highlight that current unlearning benchmarks focus too much on direct memorization rather than the broader risk of models reconstructing private data indirectly. This suggests that privacy frameworks need to move beyond one-time deletions and instead focus on continuous privacy monitoring to prevent unintended leaks.

A more robust approach to AI privacy would integrate unlearning into the entire model lifecycle rather than treating it as an afterthought. This could include adversarial probes that test whether a model can still infer sensitive data after an unlearning attempt and a two-tiered privacy evaluation system—one checking for direct memorization and another assessing whether the model can still generate private information through inference. These insights emphasize the need for privacy safeguards that are proactive, ensuring that AI models don't just forget but also stop making privacy-violating predictions in the first place.

4 Conclusion

Our analysis demonstrates that current privacy frameworks inadequately address the contextual nature of language in LLMs. Through examining recent research, we identified four key approaches to address these challenges: context-aware differential privacy that formalizes contextual sensitivity, contextual integrity frameworks that incorporate normative expectations, comprehensive privacy considerations for pre-training data, and robust unlearning mechanisms that address both direct memorization and inference patterns. Several open questions remain unresolved: How can we scale context-aware differential privacy to commercial LLMs while maintaining performance? How do we establish metrics that can accurately evaluate unlearning mechanisms against both direct and inference-based privacy leaks? How might different regulatory frameworks balance innovation with meaningful privacy protections across diverse social contexts? Future work should focus on developing end-to-end privacy guarantees from data collection through pre-training to downstream applications, creating domain-specific privacy benchmarks for healthcare, finance, and education, and investigating how privacy requirements should evolve as LLMs advance in capabilities and deployment scenarios.

5 Embedded Ethics Discussion

For an introductory NLP/AI course, we propose a three-module curriculum on "Privacy-Transparency Balance in LLMs" that translates our research into accessible learning experiences:

5.1 Module 1: Understanding the Privacy Paradox (Lecture + Discussion)

Students analyze real LLM privacy failures to understand (Brown et al., 2022) framework, examining how traditional privacy approaches break down due to context sensitivity and evolving language. The session concludes with students identifying privacy vulnerabilities in AI applications they use daily, demonstrating privacy's contextual nature firsthand.

5.2 Module 2: Contextual Integrity in Practice (Coding Homework)

Students implement a simplified contextual integrity framework by:

- Building a classifier that evaluates information sharing appropriateness across different domains
- Implementing basic differential privacy with adjustable privacy budgets
- Analyzing precision-utility tradeoffs in various social contexts

5.3 Module 3: Auditing for Privacy Leakage (Capstone Project)

Students develop privacy auditing tools by:

- Creating adversarial prompts to test commercial LLM APIs for information leakage
- Building a domain-specific privacy benchmark (healthcare, finance, or education)
- Proposing policy recommendations based on empirical findings

This curriculum bridges theory and practice, producing real privacy evaluation tools while demonstrating that ethical AI requires both technical solutions and normative frameworks. The hands-on approach ensures ethics becomes intrinsic to AI development rather than an afterthought.

6 Contribution Statement

We all divided the work among us and contributed equally to this project.

Task	People
Project organization	Vidhi
Finalizing Idea and Project Flow	Fred, Vidhi
Finding Related Works	Fred, Vidhi
Paper 1 Analysis and Reporting	Fred
Paper 2 Analysis and Reporting	Michael
Paper 3 Analysis and Reporting	Salman
Paper 4 Analysis and Reporting	Vidhi
Latex Editing	Fred,Vidhi, Salman, Michael
Introduction	Fred
Embedded Ethics Discussion	Vidhi, Salman
Conclusion	Salman, Michael

References

Sebastian Benthall and Rachel Cummings. 2024. Integrating differential privacy and contextual integrity. *Preprint*, arXiv:2401.15774.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? *Preprint*, arXiv:2202.05520.

My H. Dinh and Ferdinando Fioretto. 2023. Context-aware differential privacy for language modeling. *Preprint*, arXiv:2301.12288.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

Yan Shvartzshnaider and Vasisht Duddu. 2025. Position: Contextual integrity washing for language models. *arXiv preprint arXiv:2501.19173*.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2024. Position: Llm unlearning benchmarks are weak measures of progress. *Preprint*, arXiv:2410.02879.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2024. Position: Considerations for differentially private learning with large-scale public pretraining. *Preprint*, arXiv:2212.06470.