# Pr$\epsilon\epsilon$mpt: Sanitizing Sensitive Prompts for LLMs

Amrita Roy Chowdhury[‖*], David Glukhov[†*], Divyam Anshumaan[‡*],
Prasad Chalasani[§], Nicolas Papernot[†], Somesh Jha[‡] and Mihir Bellare[¶]
[‖]University of Michigan, Ann Arbor
[†]University of Toronto and Vector Institute
[‡]University of Wisconsin-Madison
[§]Langroid Incorporated
[¶]University of California, San Diego

**Abstract**

The rise of large language models (LLMs) has introduced new privacy challenges, particularly during *inference* where sensitive information in prompts may be exposed to proprietary LLM APIs. In this paper, we address the problem of formally protecting the sensitive information contained in a prompt while maintaining response quality. To this end, first, we introduce a cryptographically inspired notion of a *prompt sanitizer* which transforms an input prompt to protect its sensitive tokens. Second, we propose Pr$\epsilon\epsilon$mpt, a novel system that implements a prompt sanitizer, focusing on the sensitive information that can be derived solely from the individual tokens. Pr$\epsilon\epsilon$mpt categorizes sensitive tokens into two types: (1) those where the LLM's response depends solely on the format (such as SSNs, credit card numbers), for which we use format-preserving encryption (FPE); and (2) those where the response depends on specific values, (such as age, salary) for which we apply metric differential privacy (mDP). Our evaluation demonstrates that Pr$\epsilon\epsilon$mpt is a practical method to achieve meaningful privacy guarantees, while maintaining high utility compared to unsanitized prompts, and outperforming prior methods.

- **Conference:** Network and Distributed System Security (NDSS) Symposium 2026
- **Date:** 23 - 27 February 2026
- **Venue:** San Diego, CA, USA
- **ISBN:** 979-8-9919276-8-0
- **DOI:** https://dx.doi.org/10.14722/ndss.2026.231277

[*]These authors contributed equally to this work.

# Preempt: Sanitizing Sensitive Prompts for LLMs

Amrita Roy Chowdhury*, David Glukhov*, Divyam Anshumaan*, Prasad Chalasani, Nicolas Papernot, Somesh Jha and Mihir Bellare

University of Michigan · University of Toronto · UC San Diego · University of Wisconsin–Madison · Vector Institute · Institut Vecteur

## Motivation

Users frequently expose *sensitive information (PII)* to 3rd party LLMs, which can be logged in a breach of privacy.

### Prior methods ✗

- **LLM-based:** *Rephrases* prompts to remove PII. Lacks formal guarantees.
- **Substitution-based:** *Substitutes* PII with other values. Requires a lookup-table for decryption.
- **Cryptography-based:** *C*omputationally expensive.
- **DP-based noising:** Sampling noise in the embedding or token space reduce utility.

## Our Method: Preempt

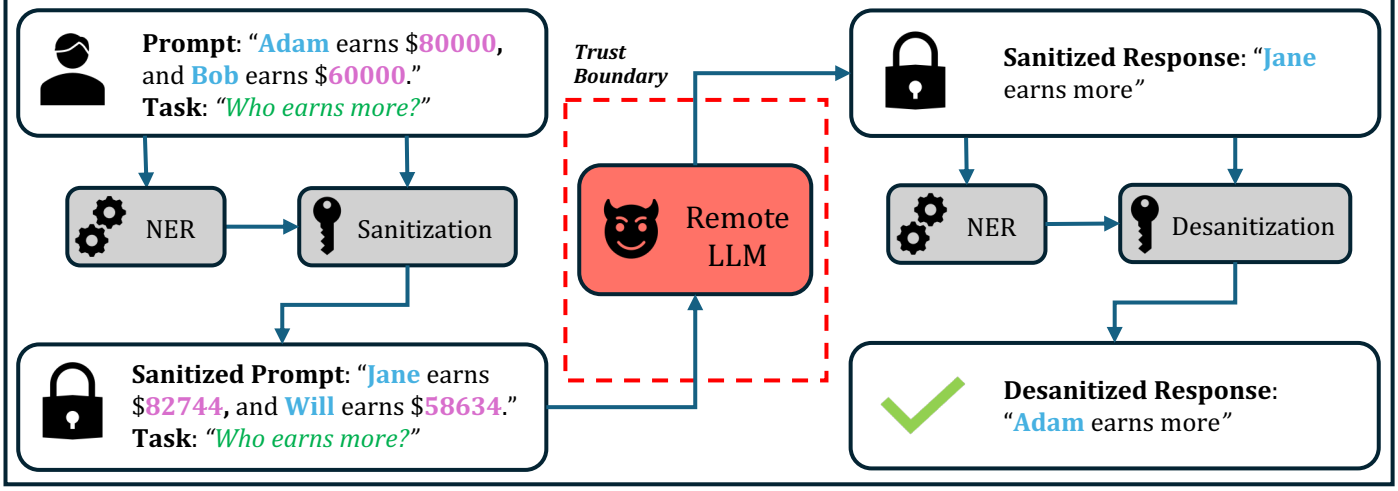
Can we *formally* protect PII in prompts while maintaining high utility and usability? Yes!

**Utility Guarantee**: Preempt targets *prompt-invariant tasks*, where LLM utility is independent of the *specific value* (Type-I) or *small variations* (Type-II) of PII attributes.

**Privacy Guarantees**:

- Type-I tokens: Protected via *format-preserving encryption*
- Type-II tokens: Protected via *metric-local differential privacy*

## Preempt - Framework




## Highlighted Results

Tasks

- ***Conv-Fin Q/A:*** A sequence of conversational numeric Q/A based on financial reports.
- ***Long-context Q/A:*** A reading comprehension Q/A task, based on summaries of books and movies.
- ***Translation:*** Translating the given English text into another language (French/Geman).
- ***RAG:*** Document retrieval.

**Preempt maintains performance compared to unsanitized text**

***Translation:*** Almost identical BLEU scores
***RAG:*** 100% accuracy
***Long-context Q/A:*** 0.934 similarity score

**Preempt maintains consistent performance *across* privacy budgets.**

***Translation:*** BLEU scores improve with increasing budget
***Conv-Fin Q/A:*** Relative error exponentially decays with budget

**Preempt outperforms contemporary LLM-based methods (PAPILLON)**

***Translation:*** Significantly outperforms PAPILLON in most cases.
***Long-context Q/A:*** Somewhat outperforms PAPILLON