LLMs AS LINGUISTIC CHAMELEONS: DECOUPLING SEMANTICS AND STRUCTURE FOR PRIVACY-PRESERVING COMMUNICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As Large Language Model (LLM) APIs become increasingly integrated into privacy-sensitive workflows, ensuring inference-time privacy without compromising task utility remains a major challenge. Existing approaches preserve most of the original semantic content to maintain downstream performance, but this also leaves exploitable cues for reconstructing the original text. This work investigates semantic decoupling, which replaces original semantics with alternative content while preserving the structure needed for LLM reasoning. Based on this idea, we propose CROSS-MAP, a bidirectional framework that maps private inputs into a different semantic domain before inference and recovers the corresponding outputs afterward. Local models are trained with multi-objective optimization to maximize semantic divergence in the mapping stage while minimizing semantic inconsistency in the recovery stage. Experiments show that CROSS-MAP reduces reconstruction success across multiple attack settings while outperforming existing baselines in utility. Code is available at `https://anonymous.4open.science/r/CROSS-MAP-7B2C/`.

1 INTRODUCTION

As LLM APIs are increasingly integrated into private workflows, a growing amount of sensitive communication, from user queries to inter-agent instructions, is sent to third-party providers. Protecting textual information has therefore become a critical challenge in the era of LLMs (Hong et al., 2024; Kan et al., 2023). A standard view in linguistics and natural language processing (NLP) is that text decomposes into semantics (meaning-bearing elements such as entities and attributes) and structure (the syntactic organization of these elements, such as word order and dependency relations) (Bloom & Lahey, 1979; Hill et al., 2025). Many reasoning tasks in LLMs are largely driven by structural patterns, implying that logical relations among entities remain valid regardless of changes in semantic content (Kim & Linzen, 2020; Li et al., 2023a). Since sensitive information is encoded primarily in semantic content rather than abstract structural templates, privacy leakage typically arises from semantics (Frikha et al., 2025; Staab et al., 2024), while structure remains largely benign.

Existing privacy-preserving approaches for LLM interaction face a fundamental trade-off between privacy and utility. To maintain downstream task performance, these methods typically keep the transformed text semantically close to the original input, since inference is performed directly on the transmitted text. Methods such as differential privacy and anonymization obfuscate sensitive information by perturbing semantic content, often degrading utility for tasks that rely on fine-grained details (Wu et al., 2024c; Hong et al., 2024; Staab et al., 2025). Other approaches introduce recovery mechanisms to mitigate semantic drift (Chowdhury et al., 2025; Kan et al., 2023; Shen et al., 2024; Chen et al., 2023b). However, since these methods operate in a semantic preserving regime, the semantic content is subject to only limited changes to retain utility, leaving exploitable cues that allow attackers to infer or reconstruct the original text (Tong et al., 2025a; Pang et al., 2025). For example, even after transformation, an LLM can still infer a user's sensitive financial and medical circumstances from a statement such as *Alice transferred* $\$50,000$ *from her savings account to pay medical debt after being diagnosed with serious disease*.

To address the dilemma inherent in semantic preserving approaches, we explore a new scenario, **semantic decoupling**, that allows LLMs to reason over preserved structural patterns while sensitive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

semantics remain local. Instead of transmitting sensitive content directly, the input is mapped into an **insensitive semantic domain** that preserves the structure but replaces the meaning. This allows the LLM to operate on a structure-preserving representation, where the original semantics is completely hidden and can only be restored locally afterward. For example, *the thief robs the bank*, *the waiter wipes the table*, and *the programmer tests the software* share the same *noun + verb + determiner + noun* structure despite expressing diverse meanings. Thus, for queries such as identifying who performs which action, the LLM can reason over the structure, while only an authorized receiver with the mapping codebook (e.g., *thief* $\leftrightarrow$ *waiter, rob* $\leftrightarrow$ *wipe, bank* $\leftrightarrow$ *table*) can recover the original text, leaving attackers observing the mapped text with little information about the sensitive content.

Historically, this paradigm has been difficult to realize. Semantic decoupling requires discovering mappings across semantic domains that preserve structural reasoning while enabling reliable recovery, which in effect requires constructing a codebook that aligns concepts across domains. Identifying such mappings demands broad linguistic and conceptual knowledge, which was largely infeasible before the emergence of modern foundation models. Recent LLMs show a strong ability to model cross-domain semantic relationships, making automated mapping increasingly plausible. However, for privacy, this capability must run locally rather than through a third-party API, since sending sensitive inputs externally would defeat the purpose of semantic decoupling. Meanwhile, general foundation models are often too large for practical local deployment and are not trained for semantic-decoupling mapping and recovery. A practical solution must therefore meet three requirements simultaneously: efficient local deployment, strict locality to avoid external exposure of private inputs, and adaptability to semantic-decoupling tasks.

This paper explores the extent to which semantic decoupling, which replaces sensitive semantic content while preserving structural reasoning patterns, can move beyond the privacy–utility dilemma inherent in semantics-preserving approaches. Specifically, we first formulate the semantics-preserving and semantic-decoupling paradigms and analyze their respective advantages from an information-theoretic perspective, characterizing when semantic decoupling offers a better privacy–utility trade-off. Second, we propose CROSS-MAP, a compact local framework for semantic-decoupling text mapping and recovery. With training objectives that encourage large semantic shifts while maintaining consistent recovery, compact local models such as Qwen-2.5-7B can be adapted to perform semantic decoupling through multi-objective optimization, achieving both stronger privacy and higher utility than semantics-preserving baselines. Finally, in addition to the defense-centric view, we adopt an attack-centric view of security by introducing an optimization-based attack to evaluate the robustness of semantic-decoupling systems against white-box adversaries, including scenarios with partial codebook leakage.

2 Related Work

This section provides a brief overview of related work. A more detailed review is provided in Appendix A.

2.1 Differential Privacy

Differential Privacy (DP) approaches for LLMs inject noise into data, gradients, or generation processes to avoid directly exposing the original private dataset to an external model (Abadi et al., 2016; Wu et al., 2024a; Yue et al., 2023; Wang et al., 2025; Wu et al., 2024c; Utpala et al., 2023; Flemings & Annavaram, 2024; Hong et al., 2024). However, DP noise can significantly degrade model performance. To preserve downstream utility, existing methods often either restrict DP protection to less critical components (Chowdhury et al., 2025; Tong et al., 2025b) or focus on tasks that are relatively insensitive to fine-grained details, such as sentiment analysis (Mattern et al., 2022; Kurakin et al., 2023).

2.2 Text Anonymization and Sanitization

Text anonymization aims to transform text to reduce the risk of revealing personal information (e.g., identity or sensitive attributes) while preserving downstream utility (Pilán et al., 2022). Prior work demonstrates that modern LLMs have the potential to both anonymize and deanonymize text, highlighting risks of malicious re-identification attacks (Patsakis & Lykousas, 2023; Staab et al., 2024; 2025). Text sanitization extends beyond protecting personal identity information to broader

sensitive content (Yue et al., 2021; Li et al., 2025; Chen et al., 2023a). To further improve downstream utility, some studies introduce an explicit desanitization step to form a bidirectional framework (Kan et al., 2023; Shen et al., 2024; Chen et al., 2023b; Chowdhury et al., 2025). This line of works shares a consensus that the mapped text and original text must be within the same semantic space for more accurate LLM inference.

3 PRELIMINARIES

This section presents necessary knowledge for understanding the difference between the proposed method and related work.

3.1 PROBLEM FORMULATION

A framework that contains both mapping and recovery processes is termed a **bidirectional privacy-preserving framework** in this work. Let $\mathcal{X}$ be the space of original texts and $\mathcal{Y}$ the space of mapped texts. A mapping process $\mathcal{M}$ produces a mapped output $y \sim \mathcal{M}(x)$ from the original text x. A recovery process $\mathcal{R}$ produces a recovered text $\tilde{x} \sim \mathcal{R}(y)$ from y. Privacy-preserving communication with LLMs aims to maintain the downstream task utility while protecting the original input x. Without protection, the LLM outputs a model response $x_{\mathrm{a}} = \mathcal{LLM}(x)$ upon receiving the input x. But to prevent the LLM from directly observing the original input x, the mapping mechanism $\mathcal{M}$ first conducts the mapping $x \mapsto y$ and then sends $y \in \mathcal{Y}$ to the LLM. Given the mapped input y, the LLM outputs a model response $y_{\mathrm{a}} = \mathcal{LLM}(y)$.

After receiving the model response y_{a} from the LLM, the recovery mechanism $\mathcal{R}$ produces a recovered response $\tilde{x}_{\mathrm{a}} = \mathcal{R}(y_{\mathrm{a}})$.The downstream task utility requires the recovered response $\tilde{x}_{\mathrm{a}}$ to be consistent with the ground-truth response x_{a}:

$$\mathrm{Sim}(x_{\mathrm{a}}, \tilde{x}_{\mathrm{a}}) \geq 1 - \epsilon,$$

where $\mathrm{Sim}(\cdot, \cdot)$ is a text similarity measure, e.g., embedding cosine, BERTScore, or an LLM-judge. The threshold ϵ is a measure for the drift in the model response induced by the mapping and recovery processes.

3.2 TWO MAPPING PARADIGMS

According to linguistic research, language is comprised of three primary components: form, content, and use (Bloom & Lahey, 1979; Hill et al., 2025). Form covers syntax, morphology, and orthography. Content denotes semantics, while use refers to pragmatics, which relates to the context of communication. Consequently, any written text x itself can be decomposed into structural component x_c (e.g., sentence organization, word order, and relationships between words) and semantic element x_s (e.g., entities, phrases, and key spans).

Under such decomposition, we present the key difference between existing mapping mechanisms, which is summarized as *semantic-preserving mapping*, and the proposed *semantic-decoupling mapping*: The goal of *semantic-preserving mapping* can be formulated as solving the following optimization problem

$$\max_{\mathcal{M},\mathcal{R}} \mathrm{Sim}(\mathcal{LLM}(x),\ \mathcal{R}(\mathcal{LLM}(\mathcal{M}(x)))),$$

subject to the *semantic similarity constraint*:

$$\mathrm{Sim}(x_s,\ \mathcal{M}(x_s)) \geq 1 - \epsilon_s.$$

In contrast, the goal of *semantic-decoupling mapping* can be formulated as subject to the *structural similarity constraint*:

$$\mathrm{Sim}(x_c,\ \mathcal{M}(x_c)) \geq 1 - \epsilon_c.$$

Briefly speaking, both mapping paradigms aim to maximize the similarity between the ground-truth LLM response $x_{\mathrm{a}} = \mathcal{LLM}(x)$ and the recovered response $\tilde{x}_a = \mathcal{R}(\mathcal{LLM}(\mathcal{M}(x)))$. *Semantic-preserving mapping* requires the mapped text $\mathcal{M}(x)$ to remain semantically similar to the original text x. For example, *the number increases rapidly* $\mapsto$ *the value rises quickly*. However, the *semantic-decoupling mapping* removes this constraint on semantic similarity. Instead, it requires structure

Symbol	Meaning
x, y, z	original text, mapped text, and LLM response
x_s, x_c	semantic and structural components of x
λ_s, λ_c	task dependence on semantic and structural information, respectively
$\mathcal{M}_{\mathrm{SD}}, \mathcal{M}_{\mathrm{SP}}$	semantic-decoupling /semantic-preserving mapping mechanisms
$\Delta\rho_s, \Delta\rho_c$	how much less semantic and structural information from the original text is preserved by $\mathcal{M}_{\mathrm{SD}}$ than by $\mathcal{M}_{\mathrm{SP}}$ in the mapped text (i.e., $\mathcal{M}_{\mathrm{SD}}$'s compression gain)
$\Delta\eta_s, \Delta\eta_c$	how much less task-relevant semantic and structural information is preserved by $\mathcal{M}_{\mathrm{SD}}$ than by $\mathcal{M}_{\mathrm{SP}}$ in the mapped text (i.e., $\mathcal{M}_{\mathrm{SD}}$'s task-relevant information loss)

Table 1: Core notation used in the main text. Other notations used in the proof are defined in Appendix B.

invariance between the original text x and the mapped text $\mathcal{M}(x)$. For example, *The number increases rapidly* $\mapsto$ *The rain drops heavily*.

4 An Information-Theoretic View

The difference between *semantic-preserving* and *semantic-decoupling* mappings can be understood from an information-theoretic perspective. Recall the mapping pipeline $(x_c, x_s) \xrightarrow{\mathcal{M}} (y_c, y_s) \xrightarrow{\mathcal{LLM}} z$, where x_c and x_s are the structural and semantic components of the original text x, y_c and y_s are the structural and semantic components of the mapped text y, and z is the LLM response on y.

Information bottleneck objective. Searching for the optimal mapping mechanism $\mathcal{M}$ is equivalent to solving the following optimization problem:

$$\mathcal{G}(\mathcal{M}) = \min_{\mathcal{M}} I(x; y) - \beta I(y; z), \tag{1}$$

where $I(x; y)$ quantifies the amount of information from the original text retained within the mapped text, and $I(y; z)$ measures the task-relevant information preserved in the mapped text for the downstream LLM. For readability, Table 1 summarizes the core notation used in the main text.

Theorem 1. *Let $\mathcal{M}_{\mathrm{SP}}$ and $\mathcal{M}_{\mathrm{SD}}$ denote semantic-preserving and semantic-decoupling mappings respectively under the same task distribution over inputs x. Assume semantic and structural information is independent in any text. Then, the information bottleneck objective satisfies*

$$\mathcal{G}(\mathcal{M}_{\mathrm{SD}}) < \mathcal{G}(\mathcal{M}_{\mathrm{SP}}),$$

i.e., $\mathcal{M}_{\mathrm{SD}}$ achieves a lower information bottleneck objective than $\mathcal{M}_{\mathrm{SP}}$, if the downstream task's semantic-dependence coefficient λ_s is smaller than this threshold:

$$\lambda_s < \frac{r \cdot \Delta\rho_s + (1 - r)\Delta\rho_c - \beta' \cdot \Delta\eta_c}{\beta'(\Delta\eta_s - \Delta\eta_c)}, \tag{2}$$

where β' and r are some constants. The detailed definitions of involved quantities and a full derivation are provided in Appendix B.

Corollary 1. *$\mathcal{M}_{\mathrm{SD}}$ achieves a lower information bottleneck objective than $\mathcal{M}_{\mathrm{SP}}$, if its gains in compression outweigh its task-relevant information loss in the following way:*

$$r \cdot \Delta\rho_s + (1 - r)\Delta\rho_c > \beta'(\Delta\eta_s \cdot \lambda_s + \Delta\eta_c \cdot \lambda_c).$$

Remark 1. *Eq. (2) formalizes an intuitive trade-off that $\mathcal{M}_{\mathrm{SD}}$ is advantageous when it brings substantial compression gains while only mildly harming the task-relevant information needed by the downstream task: In the numerator, larger $\Delta\rho_s$ or $\Delta\rho_c$ means that $\mathcal{M}_{\mathrm{SD}}$ removes more semantic or structural information from the original text than $\mathcal{M}_{\mathrm{SP}}$, leading to a stronger compression and privacy advantage. This enlarges the threshold, so $\mathcal{M}_{\mathrm{SD}}$ remains preferable for a wider range of tasks. The weight r reflects where most of the entropy in the original input lies. When r is larger, the information carried by the original text is dominated by its semantic content.*

By contrast, larger $\Delta\eta_s$ means that semantic decoupling discards more task-relevant semantic information, which increases the denominator and thus lowers the threshold. Larger $\Delta\eta_c$ has a similar effect through the term $-\beta'\Delta\eta_c$ in the numerator. Finally, a larger β' implies that downstream utility is weighted more heavily relative to compression, so the threshold becomes more stringent.

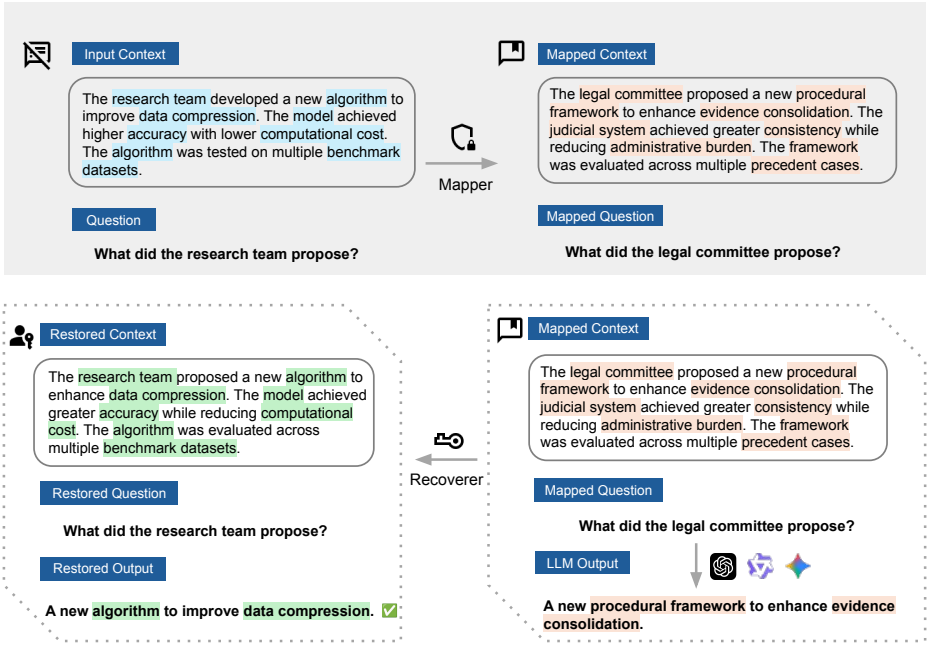


Figure 1: Illustration of inference-time protection with CROSS-MAP.

Privacy benefits with compression. According to the variational information bottleneck (VIB) framework, the mutual information terms in Equation (1) can be bounded as follows:

$$I(x;y) = \mathbb{E}_{y\sim p(y)}\left[KL(p(x|y) \parallel p(x))\right],$$

and

$$I(y;z) \geq \mathbb{E}_{y,z}\left[\log q(z|y)\right],$$

where $q(z|y)$ is an auxiliary variational distribution used to approximate the posterior $p(z|y)$. On one hand, $\mathcal{M}_{\mathrm{SD}}$ reduces $I(x;y)$ by decoupling the semantic information of x and y, which encourages smaller KL divergence between $p(y|x)$ and $p(y)$. Equivalently, the attacker posterior $p(x|y)$ becomes closer to the true prior $p(x)$ on average, implying that the mapped output y reveals less usable information about the original text x. Therefore, this compression directly enhances privacy against attackers observing y. [1]

On the other hand, such privacy benefits of $\mathcal{M}_{\mathrm{SD}}$ come at the cost of dropping more task-relevant semantic information compared to $\mathcal{M}_{\mathrm{SP}}$, which might impact the $q(z|y)$ modelled by the downstream LLM. As implied by Equation (2), if the downstream task's dependence on semantic information can tolerate such loss, $\mathcal{M}_{\mathrm{SD}}$ dominates $\mathcal{M}_{\mathrm{SP}}$ with compression benefits outweighing the task information loss.

5 CROSS-MAP

This section presents CROSS-MAP, a semantic-decoupling framework designed for cross-domain mapping and recovery.

5.1 Key Components

To address the challenges of highly-unstructured text inputs, two key designs are needed: (i) a structured mapping and a recovery module, and (ii) a verifiable reward and metric design.

Structured mapping and recovery. Given an original input x, the mapper $\mathcal{M}_\phi$ produces a cross-domain representation $(\tau, D, y) \sim \mathcal{M}_\phi(x)$ consisting of a target-domain theme τ, a dictionary D, and a mapped text y. Here $D = \{(s_i, \hat{s}_i)\}_{i=1}^{m}$ maps each original span s_i to a target-domain span $\hat{s}_i$.

[1] An in-depth analysis of the connection between KL divergence and posterior-based attack success rate is provided in Appendix C.

The recoverer $\mathcal{R}_\psi$ reconstructs the restored text $\tilde{x} \sim \mathcal{R}_\psi(\tau, D, y)$ from (τ, D, y). Both $\mathcal{M}_\phi$ and $\mathcal{R}_\psi$ are implemented as instruction-following LLMs with structured outputs (e.g., JSON).

Verifiable reward. A compact set of score metrics is needed to guide the training of the mapper and the recoverer:

- **Dictionary coverage** $c_{\text{dict}}(x, D)$**.** Let $x = (x_1, \ldots, x_n)$ be the token sequence of length n. Each original span s_i in D corresponds to a contiguous index interval $I_i \subseteq \{1, \ldots, n\}$ in x. Define the set of covered token positions as $\mathcal{I}(D) := \bigcup_{i=1}^{m} I_i$. The dictionary coverage is the fraction of tokens in x that are covered by a dictionary span: $c_{\text{dict}}(x, D) = |\mathcal{I}(D)|/n$. Intuitively, higher c_{dict} means a larger portion of the original text x is explicitly handled by the mapper through D.
- **Dictionary distance** $d_{\text{dict}}(D)$**.** Using a cosine similarity measure $\text{Sim}(\cdot, \cdot)$ that takes in span embeddings, the distance between spans in D is measured as $d_{\text{dict}}(D) = 1 - \frac{1}{m}\sum_{i=1}^{m} \text{Sim}(s_i, \hat{s}_i)$.
- **Text distance** $d_{\text{text}}(x, y)$**.** Using a cosine similarity measure that takes in sentence embeddings, the text distance between x and y is measured as $d_{\text{text}}(x, y) = 1 - \frac{1}{L}\sum_{j=1}^{L} \text{Sim}(x_j, y_j)$, where $\{x_j\}_{j=1}^{L}$ and $\{y_j\}_{j=1}^{L}$ are sentence splits of x and y.
- **Fluency** $f_{\text{text}}(y)$**.** Let $(y_1, \ldots, y_T)$ be the token sequence of the mapped text y. The fluency score is set to the negative log-perplexity $-\frac{1}{T}\sum_{t=1}^{T} \log p_{\text{LM}}\big(y_t | y_{<t}\big)$, where p_{LM} represents the output distribution of a reference language model.

Finally, the above metrics are aggregated into a single scalar reward:

$$\begin{aligned} r(x, \tau, D, y, \tilde{x}) = {} & w_1\, c_{\text{dict}}(x, D) + w_2\, d_{\text{dict}}(D) \\ & + w_3\, d_{\text{text}}(x, y) + w_4\, f_{\text{text}}(y) \\ & + w_5\, (1 - d_{\text{text}}(x, \tilde{x})), \end{aligned}$$

with nonnegative weights $\{w_1, w_2, w_3, w_4, w_5\}$. The last term measures the consistency between the original text x and recovered text $\tilde{x}$

5.2 Training: SFT to DPO

The training pipeline has three stages: (i) seed data synthesis, (ii) supervised fine-tuning (SFT) for structured mapping/recovery, and (iii) direct preference optimization (DPO) for multi-objective alignment.

Stage 1: Data synthesis. We first use GPT-5.2 to synthesize a small subset[2] of training instances $\mathcal{D}_{\text{synth}} = \{(x \xrightarrow{\mathcal{M}} (\tau, D, y)),\ ((\tau, D, y) \xrightarrow{\mathcal{R}} \tilde{x})\}$.

Stage 2: SFT. The goal is to train two instruction-following models: a private mapper $\mathcal{M}_\phi$ and a recoverer $\mathcal{R}_\psi$. Both are initialized from the same base LLM and fine-tuned with parameter-efficient adapters. The mapper learns to generate (τ, D, y) from x, while the recoverer learns to generate $\tilde{x}$ from (τ, D, y).

Stage 3: DPO. SFT establishes a strong base, but it does not directly optimize the multi-objective reward. DPO is therefore employed to refine the mapper with preference pairs obtained by sampling and scoring. During preference construction, for each x, we sample K mapped candidates $(\tau^{(k)}, D^{(k)}, y^{(k)}) \sim p_\phi(\cdot \mid x), \quad k = 1, \ldots, K$, and obtain a recovery for each candidate as $\tilde{x}^{(k)} \sim p_\psi(\cdot \mid \tau^{(k)}, D^{(k)}, y^{(k)})$. The scalar reward is computed as $r^{(k)} := r(x, \tau^{(k)}, D^{(k)}, y^{(k)}, \tilde{x}^{(k)})$, and a preference pair (k^+, k^-) is selected such that $r^{(k^+)} \geq r^{(k^-)} + \gamma$ for a margin $\gamma > 0$. Let $o^+ := o^{(k^+)}$ and $o^- := o^{(k^-)}$ denote the preferred and dispreferred mapper outputs, and p_{ϕ_0} be the frozen reference model (typically the SFT checkpoint). DPO optimizes the mapper parameters ϕ by increasing the relative likelihood of high-reward mapped outputs over low-reward ones.

[2]These instances are available in the provided GitHub repository.

5.3 WORKFLOW WITH LLMS AS THIRD-PARTY API

Figure 1 illustrates the workflow of the proposed CROSS-MAP with the trained mapper and recoverer. After training, the private mapper $\mathcal{M}_\phi$ and the private recoverer $\mathcal{R}_\psi$ are deployed locally (e.g., on the user side), while the downstream LLM is accessed as a third-party API. Before the transmission, $\mathcal{M}_\phi$ first maps the original input x to a mapped version y. The LLM then produces a response y_a and transmits it back to the recoverer. The private recoverer can subsequently recover the original-domain answer $\tilde{x}_\mathrm{a}$ by reasoning jointly over the dictionary and the LLM response y_a. Therefore, users can interact with external APIs or proprietary services under the protection of the local mapper and recoverer via CROSS-MAP. It allows the external model to operate on an alternative semantic domain while keeping the original sensitive content local. Algorithm 1 in Appendix D summarizes this workflow.

5.4 ADVERSARIAL ATTACKS

To evaluate the security of CROSS-MAP, we simulate a white-box attacking scenario where the attacker has access to the mapper $\mathcal{M}_\phi$ and even to a partially leaked dictionary. Assume the attacker observes a mapped output y and attempts to recover the original input x using a parametric attack policy $p_\theta(\cdot|y)$.

Attack objective and reward. Given a candidate reconstruction $\hat{x} \sim p_\theta(\cdot|y)$, the attacker can evaluate how well it explains the observation by passing $\hat{x}$ through the mapper to obtain $\hat{o} := (\hat{\tau}, \hat{D}, \hat{y}) \sim \mathcal{M}_\phi(\hat{x})$. A successful $\hat{x}$ should be mapped to $\hat{y}$ that is close to the observed y. The attack reward is therefore defined as the similarity between $\hat{y}$ and y: $\hat{r}(\hat{x}; y) = 1 - d_{\mathrm{text}}(\hat{y}, y)$. The attacker aims to find a reconstruction $\hat{x}$ that maximizes $\hat{r}(\hat{x}; y)$.

DPO-based attacker. An optimization-based attacker can align $p_\theta(\cdot|y)$ to the reward $\hat{r}(\hat{x}; y)$ via DPO. For each observed y, multiple candidate reconstructions $\{\hat{x}^{(k)}\}_{k=1}^{K}$ are sampled from $p_\theta(\cdot|y)$, then mapped by $\mathcal{M}_\phi$ to obtain $\hat{y}^{(k)}$, and finally scored by $\hat{r}^{(k)} = 1 - d_{\mathrm{text}}(\hat{y}^{(k)}, y)$. A preference pair $(\hat{x}^+, \hat{x}^-)$ is then constructed by selecting two candidates with a reward gap $\hat{r}(\hat{x}^+; y) \geq \hat{r}(\hat{x}^-; y) + \gamma_{\mathrm{atk}}$ for a margin $\gamma_{\mathrm{atk}} > 0$. A detailed theoretical analysis of the attacker in the presence of dictionary leakage is provided in Appendix E.

6 EXPERIMENTS

CROSS-MAP is evaluated along two dimensions: (i) utility on question answering (QA), summarization, and generation tasks; and (ii) security against de-anonymization, black-box span restoration, and optimization-based white-box attacks. Experiments are conducted on datasets from multiple semantic domains, including medicine, finance, news, and daily life. A full description of the experimental setup, including the datasets and models used, is provided in Appendix F.1.

6.1 UTILITY AND PRIVACY

Table 2 reports the utility and privacy of multiple methods across four benchmarks. Across all datasets, CROSS-MAP consistently achieves the strongest mapped-text privacy while largely preserving downstream QA utility relative to other baselines. For example, on SQuAD, the distance between mapped text and original text increases from HAS's 0.238 to 0.752 while QA accuracy increases from HAS's 91.84 to 92.98. Similar trends are observed on other datasets. Plain QA and Pr$\varepsilon\varepsilon$mpt retain the highest downstream utility but provide negligible or weak privacy. For Pr$\varepsilon\varepsilon$mpt, the mapped-text privacy scores remain low overall, reflecting its minimal perturbation on the original text. InferDPT delivers only partial privacy gains while causing a substantial utility collapse (e.g., on SQuAD, F1 drops from 84.10 to 74.04), highlighting that perturbation without an explicit recovery mechanism distorts answer-critical semantics. As a bidirectional framework, HAS attains higher utility than InferDPT, but its privacy remains below CROSS-MAP, suggesting that residual semantic cues are still preserved in the mapped text and thus remain exposable to external LLMs. The consistent privacy gains and the near-preserved downstream utility achieved by CROSS-MAP demonstrate that a bidirectional framework with semantic decoupling yields the best privacy-utility trade-off among the compared methods. Table 16 in Appendix F.2 provides examples of different mapping methods.

Table 2: Privacy and utility results across datasets for Plain QA, Prεεmpt (Chowdhury et al., 2025), InferDPT (Tong et al., 2025b), HAS (Chen et al., 2023b), and CROSS-MAP. The results are reported using Qwen2.5-14B as the base model for both the mapper and the recoverer in CROSS-MAP. An LLAMA-8B model is used for downstream QA. Please refer to Appendix F.2 for detailed definitions of the metrics.

		Downstream QA utility			**Recovered text utility**				**Mapped text privacy**		
Dataset	Method	EM↑	F1↑	ACC↑	Flu. Loss↓	Sim.↑	BLEU↑	R-1↑	c_{dict} ↑	d_{dict} ↑	d_{text} ↑
SQuAD	Plain QA	76.42	84.10	94.73	4.202	1.000	1.000	1.000	0.000	0.000	0.000
	Prεεmpt	75.78	84.02	94.70	4.365	1.000	1.000	1.000	0.006	0.142	0.006
	InferDPT	68.21	74.04	81.34	5.597	0.737	0.520	0.710	0.172	0.289	0.263
	HAS	72.97	82.10	91.84	4.212	0.973	0.939	0.975	0.181	0.408	0.238
	CROSS-MAP	**74.94**	**83.22**	**92.98**	4.703	0.826	0.586	0.800	**0.413**	**0.753**	**0.752**
BioASQ	Plain QA	13.48	32.02	96.84	4.313	1.000	1.000	1.000	0.000	0.000	0.000
	Prεεmpt	13.44	31.99	96.81	4.156	1.000	1.000	1.000	0.041	0.497	0.151
	InferDPT	8.97	19.88	77.44	4.117	0.709	0.540	0.730	0.094	0.375	0.291
	HAS	11.91	28.45	95.21	4.346	0.998	0.958	0.977	0.103	0.553	0.203
	CROSS-MAP	**12.73**	**30.90**	**96.10**	5.128	0.879	0.610	0.805	**0.604**	**0.858**	**0.919**
FinCausal	Plain QA	59.27	77.61	81.20	4.712	1.000	1.000	1.000	0.000	0.000	0.000
	Prεεmpt	59.20	77.54	81.15	4.722	1.000	1.000	1.000	0.016	0.150	0.037
	InferDPT	58.63	77.10	79.06	4.824	0.964	0.893	0.949	0.054	0.238	0.036
	HAS	57.40	75.89	79.24	4.645	0.968	0.921	0.956	0.092	0.294	0.139
	CROSS-MAP	**58.99**	**77.33**	**80.61**	4.970	0.860	0.733	0.877	**0.540**	**0.710**	**0.678**
NarrativeQA	Plain QA	21.12	44.21	90.68	4.518	1.000	1.000	1.000	0.000	0.000	0.000
	Prεεmpt	21.07	44.15	90.65	4.681	1.000	1.000	1.000	0.011	0.516	0.117
	InferDPT	15.20	38.74	72.12	4.545	0.846	0.910	0.955	0.014	0.251	0.154
	HAS	19.61	41.92	87.80	4.512	0.943	0.897	0.954	0.094	0.541	0.341
	CROSS-MAP	**20.49**	**43.22**	**89.67**	5.138	0.829	0.689	0.863	**0.313**	**0.796**	**0.880**

To broaden the evaluation beyond QA, the SQuAD-trained mapper and recoverer are also evaluated on two non-QA task families without extra in-domain training. Table 3 in Appendix F.2 summarizes the results. On CNN/DailyMail summarization, CROSS-MAP preserves summary quality, yielding a BERTScore of 0.877 compared to 0.882 for direct Plain LLM inference, while reducing the sensitive-span retention from 100.0% to 7.3%. On CommonGen open-ended generation, CROSS-MAP preserves high concept coverage (0.891) while reducing retention from 98.3% to 18.9%. These results support Theorem 1: semantic decoupling is strongest when the task-relevant structure is sufficient, but it can still provide a superior privacy-utility trade-off for semantic-dependent generation when a good mapper performs domain-level dictionary construction and the recoverer performs soft semantic recovery instead of string replacement, which helps raise the threshold and broaden the applicable scope.

6.2 Security Against Attacks

This subsection reports and compares the security of different methods in three attack scenarios.

6.2.1 De-anonymization Attacks

LLM attribute inference is a strong way for de-anonymization. The goal is to use an LLM to infer private attributes (e.g., gender, age, and location) from the mapped text. This subsection reports the success rate of LLM attribute inference on SynthPAI (Yukhymenko et al., 2024) using GPT-4 as the attacker as in GPT-AA (Staab et al., 2025). Figure 5 in Appendix F.2 shows that CROSS-MAP yields the lowest inference accuracy and certainty among all compared methods. The concrete numbers for each method are reported in Table 5 in Appendix F.2.

6.2.2 Black-box Span Restoration Attacks

Span restoration attacks aim to reconstruct the exact protected spans from the mapped text. The Attack Success Rate (ASR) is defined as the span restoration accuracy, calculated as the ratio of correctly restored spans to the total number of protected spans. Figure 2 visualizes the trend of span restoration across datasets as k varies. Plain QA consistently yields the highest ASR, while

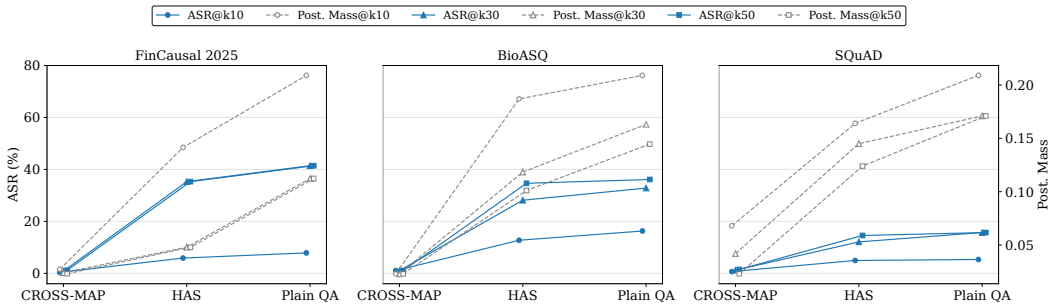


Figure 2: Correlation between span restoration accuracy and posterior mass on protected spans.

HAS offers only limited improvement. In contrast, CROSS-MAP substantially reduces restoration success across all datasets and all values of k. This suggests that preserving semantic anchors makes posterior-based restoration easier for the attacker. Table 17 in Appendix F.2 also reports the posterior mass on the ground-truth span, as defined in Appendix C. Across datasets and different values of k, methods with lower posterior mass exhibit lower ASR. For instance, on FinCausal 2025 at $k = 50$, CROSS-MAP assigns the lowest posterior mass (0.002) together with the lowest ASR (1.19%), whereas HAS and Plain QA show higher posterior mass (0.026 and 0.088) and correspondingly higher ASR (35.29% and 41.43%).

6.2.3 White-box Optimization-based Attacks

The attacker is initialized from the trained mapper and then optimized using DPO algorithm as described in Section 5.4. ASR here refers to the similarity score $1 - d_{\text{text}}(\hat{x}, x)$ between the restored text $\hat{x}$ and the original text x. Table 13 and Table 14 in Appendix F.2 show that restoration becomes easier as the dictionary leakage ratio ρ increases for both datasets, and that the DPO-based attacker consistently outperforms the base attacker without optimization at every ρ. Figure 6 in Appendix F.2 shows that higher leakage also helps DPO-based attacker to achieve larger gains over the base attacker. However, it is noticeable that even under this optimization-based white-box attacker, the improvement brought by optimization under CROSS-MAP remains bounded. Across both datasets and all leakage ratios ρ, the maximum relative gain of the DPO attacker over the base attacker is 29.4%, achieved on NarrativeQA at $\rho = 0.1$ (where the DPO attacker's ASR is still only 24.2%). On SQuAD, the maximum relative gain is 13.6% (achieved at $\rho = 0.1$, with the DPO attacker's ASR still only 32.6%). This bounded improvement suggests that, even when the mapper's parameters are exposed and the attacker is allowed to optimize based on extra knowledge, e.g., some pairs from the mapping dictionary, the restoration advantage that can be extracted by optimization is still limited, supporting the robustness of CROSS-MAP against optimization-driven white-box restoration.

6.3 Ablation Study

The ablation studies in this paper systematically evaluate the key factors impacting model deployment, including input complexity, paraphrase robustness, target-domain selection, dictionary coverage, training scale, external LLM capabilities, mapper/recoverer architecture, and training strategies. These findings offer empirical guidelines for optimal experimental configurations. Detailed tables and extended analyses are provided in Appendix F.2.

7 Conclusion

This paper proposes CROSS-MAP, a semantic-decoupling framework for inference-time text protection. By replacing semantic content while preserving the structural patterns needed for downstream LLM reasoning, CROSS-MAP addresses the privacy–utility dilemma inherent in semantics-preserving approaches. An information-theoretic analysis characterizes how and when semantic decoupling provides a better privacy–utility trade-off. Experiments on QA, summarization, and generation tasks validate the analysis and demonstrate that CROSS-MAP outperforms existing baselines in both privacy and utility. These findings highlight the potential of semantic decoupling as a practical direction for privacy-preserving LLM inference.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets: GPT-5.2 synthesized a subset of the mapper and recoverer training instances, as described in Section 5.2. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, implementing methods, cleaning or reformatting datasets, or interpreting results. Additionally, we used generative AI tools for editing the research paper to improve readability. We have reviewed all AI-assisted work by carefully checking the paper's wording and presentation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

A potential dual-use risk arises from semantic decoupling. By replacing sensitive semantics while preserving structural patterns, the method could also be used to obscure malicious intent from external oversight while retaining enough structure for downstream reasoning. This creates a potential safety risk: harmful requests or plans may become less interpretable to LLM service providers even when their functional structure is preserved. Although CROSS-MAP is intended for privacy protection, its practical deployment should therefore incorporate safeguards against misuse.

REPRODUCIBILITY STATEMENT

Section 5 describes the method and training pipeline. Section 6 reports the evaluations. The appendix provides proofs, attack analysis, implementation details, prompts, and additional results. An anonymous code repository is linked in the abstract.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Lois Bloom and Margaret Lahey. Language development and language disorders. *Language*, 55:945, 1979.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL*, pp. 5747–5758, 2023a.

Yu Chen, Tingxin Li, Huiming Liu, and Yang Yu. Hide and seek (HaS): A lightweight framework for prompt privacy protection. *arXiv preprint arXiv:2309.03057*, 2023b.

Zhiyu Chen, Yu Li, Suochao Zhang, Jingbo Zhou, Jiwen Zhou, Chenfu Bao, and Dianhai Yu. A framework for cost-effective and self-adaptive LLM shaking and recovery mechanism. *arXiv preprint arXiv:2403.07283*, 2024.

Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan, Prasad Chalasani, Nicolas Papernot, Somesh Jha, and Mihir Bellare. Pr$\varepsilon\varepsilon$mpt: Sanitizing sensitive prompts for LLMs. *arXiv preprint arXiv:2504.05147*, 2025.

James Flemings and Murali Annavaram. Differentially private knowledge distillation via synthetic text generation. In *Findings of the Association for Computational Linguistics: ACL*, pp. 12957–12968, 2024.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. IncogniText: Privacy-enhancing conditional text anonymization via LLM-based private attribute randomization. In *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pp. 2490–2501, 2025.

Elizabeth Hill, Kate Tonta, Mark Boyes, Courtenay Frazier Norbury, Sarah Griffiths, Shaun Goh, and Brooke Ryan. Why would someone like me with DLD want to sit in a room and talk? How would that make me feel better?!: Developmental language disorder and the language demands of cognitive behaviour therapy. *International Journal of Cognitive Behavioral Therapy*, 18:405–424, 2025.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoyang Hou, Jian Liu, Jingyu Li, Yuhan Li, Jiawen Zhang, Wen-jie Lu, Cheng Hong, and Kui Ren. CipherGPT: Secure two-party GPT inference. *IEEE Transactions on Dependable and Secure Computing*, pp. 1–16, 2026.

Sofoklis Kakouros, Juraj Šimko, Martti Vainio, and Antti Suni. Investigating the utility of surprisal from large language models for speech synthesis prosody. *arXiv preprint arXiv:2306.09814*, 2023.

Zhigang Kan, Linbo Qiao, Hao Yu, Liwen Peng, Yifu Gao, and Dongsheng Li. Protecting user privacy in remote conversational systems: A privacy-preserving framework based on text sanitization. *arXiv preprint arXiv:2306.08223*, 2023.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105, 2020.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. SLOG: A structural generalization benchmark for semantic parsing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3213–3232, 2023a.

Dacheng Li, Hongyi Wang, Rulin Shao, Han Guo, Eric Xing, and Hao Zhang. MPCFORMER: Fast, performant and private transformer inference with MPC. In *International Conference on Learning Representations (ICLR)*, 2023b.

Siyan Li, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy preservation from internet-based and local language model ensembles. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3371–3390, 2025.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schölkopf, and Mrinmaya Sachan. Differentially private language models for secure data sharing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4860–4873, 2022.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the Financial Technology and Natural Language Processing, the Financial Narrative Processing, and the Workshop on Large Language Models for Finance and Legal*, pp. 214–221, 2025.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, and Minhui Xue. Reconstruction of differentially private text sanitization via large language models. In *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pp. 1–17, 2025.

Constantinos Patsakis and Nikolaos Lykousas. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026, 2023.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, 2022.

Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Truthful text sanitization guided by inference attacks. *Applied Soft Computing*, 185:114013, 2025.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfenag Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016.

Aleksandr Romanov, Anna Kurtukova, Anastasia Fedotova, and Roman Meshcheryakov. Natural text anonymization using universal transformer with a self-attention. In *International Conference on Language Engineering and Applied Linguistics (LEA)*, pp. 22–37, 2019.

Zhili Shen, Zihang Xi, Ying He, Wei Tong, Jingyu Hua, and Sheng Zhong. The fire thief is also the keeper: Balancing usability and privacy in prompts. *arXiv preprint arXiv:2406.14318*, 2024.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. A4NT: Author attribute anonymity by adversarial training of neural machine translation. In *USENIX Security Symposium*, pp. 1633–1650, 2018.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Large language models are advanced anonymizers. In *International Conference on Learning Representations (ICLR)*, 2025.

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. On the vulnerability of text sanitization. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 5150–5164, 2025a.

Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, Tianwei Zhang, and Zhikun Zhang. InferDPT: Privacy-preserving inference for black-box large language models. *IEEE Transactions on Dependable and Secure Computing*, 22:4625–4640, 2025b.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 8442–8457, 2023.

Jianwei Wang, Chengming Shi, Junyao Yang, Haoran Li, Qianli Ma, Huiping Zhuang, Cen Chen, and Ziqian Zeng. RewardDS: Privacy-preserving fine-tuning for large language models via reward driven data synthesis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4479–4500, 2025.

Fan Wu, Huseyin A. Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and Robert Sim. Privately aligning language models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024a.

Haoqi Wu, Wenjing Fang, Yancheng Zheng, Junming Ma, Jin Tan, and Lei Wang. Ditto: Quantization-aware secure inference of transformers upon MPC. In *International Conference on Machine Learning (ICML)*, pp. 53346–53365, 2024b.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *International Conference on Learning Representations (ICLR)*, 2024c.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 28922–28941, 2025b.

Mu Yuan, Lan Zhang, and Xiang-Yang Li. Secure transformer inference protocol. *arXiv preprint arXiv:2312.00025*, 2023.

Patrick Yubeaton, Jianqiao Cambridge Mo, Karthik Garimella, Nandan Kumar Jha, Brandon Reagen, Chinmay Hegde, and Siddharth Garg. TruncFormer: Private LLM inference using only truncations. *arXiv preprint arXiv:2412.01042*, 2024.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL–IJCNLP*, pp. 3853–3866, 2021.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1321–1342, 2023.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:120735–120779, 2024.

Ruoyan Zhang, Zhongxiang Zheng, and Wankang Bao. Practical secure inference algorithm for a fine-tuned large language model based on fully homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 21:17–29, 2025.

A Full Literature Review

This section provides a comprehensive review of related work.

A.1 Cryptography

This line of work studies text encryption for privacy protection (Zhang et al., 2025; Wu et al., 2024b; Li et al., 2023b). Many systems for agent-to-agent communication also provide support for standard encryption primitives (Yuan et al., 2023). While these techniques offer lossless protection that allows complete recovery, they induce nontrivial computational overhead and additional communication latency. Even when optimizations are applied (Chen et al., 2024; Hou et al., 2026; Yubeaton et al., 2024), the cost and engineering complexity remain substantial in interactive inference pipelines, particularly in long-context and latency-sensitive settings, which limits the practicality of cryptography as a general-purpose solution for inference-time text protection.

A.2 Differential Privacy

Differential Privacy (DP) achieves privacy protection by perturbing data-dependent computations with randomness. As LLMs progress, there are growing concerns about their capacity to memorize sensitive information from their input data. Existing approaches include applying DP during model training, such as using DP variants of stochastic optimization when fine-tuning on private datasets (Abadi et al., 2016; Wu et al., 2024a). Another line of DP research focuses on training private generative models locally under DP constraints to synthesize data, which is then utilized for downstream fine-tuning or inference with larger models (Yue et al., 2023; Wang et al., 2025; Wu et al., 2024c; Utpala et al., 2023; Flemings & Annavaram, 2024; Hong et al., 2024). These DP-based approaches share the core principle of injecting noise into data, gradients, or generation processes to avoid directly exposing the original private dataset to an external model. However, DP noise can significantly degrade model performance by perturbing information essential for reasoning and prediction. To preserve downstream utility, existing methods often either restrict DP protection to less critical components (Chowdhury et al., 2025; Tong et al., 2025b), or focus on tasks that are relatively insensitive to exact factual and entity-level details, such as sentiment analysis (Mattern et al., 2022; Kurakin et al., 2023).

A.3 Text Anonymization and Sanitization

Text anonymization aims to transform text to reduce the risk of revealing personal information (e.g., identity or sensitive attributes) while preserving downstream utility (Pilán et al., 2022). Early neural rewriting approaches combined feature-guided edits, such as dictionary-based synonym substitution, with sequence-to-sequence generation for anonymization (Romanov et al., 2019; Shetty et al., 2018). Prior work demonstrates that modern LLMs have the potential to both anonymize and deanonymize text, highlighting risks of malicious re-identification attacks (Patsakis & Lykousas, 2023; Staab et al., 2024). Following research further proposes an LLM-based adversarial anonymization framework against LLM-driven re-identification (Staab et al., 2025). Related attempts include multi-objective optimization for utility-preserving anonymization (Yang et al., 2025b), misleading adversaries into predicting incorrect private attributes (Frikha et al., 2025), and span-level truthful anonymization via semantic generalization (Pilán et al., 2025).

Compared to text anonymization, text sanitization extends beyond protecting personal identity information to broader sensitive content (Yue et al., 2021; Li et al., 2025; Chen et al., 2023a). To further improve text utility for downstream tasks, some studies introduce an explicit desanitization step. Existing methods include approaches based on predefined plaintext–ciphertext mappings (Kan et al., 2023), methods that use a language model to generate span replacements and internalize transformation rules in a local model (Shen et al., 2024; Chen et al., 2023b), and designs that protect only specific sensitive types, e.g. numerical values, to preserve most of the original semantics (Chowdhury et al., 2025). This line of works share a consensus that the mapped text and original text must be within the same semantic space for more accurate LLM inference.

B PROOF OF THEOREM 1

The formal proof of Theorem 1 begins with the following definitions.

Preservation ratios ρ_s and ρ_c. These ratios measure the amount of semantic and structural information from the original text that is preserved in the mapped text y:

$$\rho_s = \frac{I(y_s; x_s)}{H(x_s)}, \qquad \rho_c = \frac{I(y_c; x_c)}{H(x_c)},$$

where $I(y_s; x_s)$ and $I(y_c; x_c)$ denote the mutual information between the semantic and structural components of the mapped text and those of the original text. $H(x_s)$ and $H(x_c)$ denote the entropy of the semantic and structural components of the original text.

Task-relevant preservation ratios η_s and η_c. These ratios quantify how much task-relevant semantic and structural information is transferred from the original text to the mapped text:

$$\eta_s = \frac{I(y_s; z)}{I(x_s; z)}, \qquad \eta_c = \frac{I(y_c; z)}{I(x_c; z)},$$

where $I(y_s; z)$ and $I(y_c; z)$ denote the mutual information between the semantic and structural components of the mapped text and the downstream task output z. $I(x_s; z)$ and $I(x_c; z)$ denote the mutual information between the semantic and structural components of the original text and the downstream task. These ratios characterize the fraction of task-relevant information preserved in the mapped text.

Task dependence coefficients λ_s and λ_c. These coefficients describe the relative importance of semantic and structural information in the original text for the downstream task:

$$\lambda_s = \frac{I(x_s; z)}{I(x; z)}, \qquad \lambda_c = \frac{I(x_c; z)}{I(x; z)},$$

where $I(x_s; z)$ and $I(x_c; z)$ denote the mutual information between the semantic and structural components of the original text and the downstream task output z, and $I(x; z)$ denotes the mutual information between the entire input text and the task output. Since

$$I(x; z) = I(x_s; z) + I(x_c; z),$$

it follows that

$$\lambda_s + \lambda_c = 1.$$

Compression gains $\Delta\rho_s$ and $\Delta\rho_c$. These quantities measure the difference in preserved information between the semantic-preserving and semantic-decoupling mappings:

$$\Delta\rho_s = \rho_s^{SP} - \rho_s^{SD}, \quad \Delta\rho_c = \rho_c^{SP} - \rho_c^{SD},$$

where ρ_s^{SP} and ρ_c^{SP} denote the preservation ratios under the semantic-preserving mapping, and ρ_s^{SD} and ρ_c^{SD} denote the preservation ratios under the semantic-decoupling mapping.

Task-relevant preservation differences $\Delta\eta_s$ and $\Delta\eta_c$. These quantities measure the difference in task-relevant information retained by the two mappings:

$$\Delta\eta_s = \eta_s^{SP} - \eta_s^{SD}, \quad \Delta\eta_c = \eta_c^{SP} - \eta_c^{SD},$$

where η_s^{SP} and η_c^{SP} denote the task-relevant preservation ratios under the semantic-preserving mapping, while η_s^{SD} and η_c^{SD} denote those under the semantic-decoupling mapping.

Proof of Theorem 1. Consider the information bottleneck objective of a mapping $\mathcal{M}$:

$$\mathcal{G}(\mathcal{M}) = I(x; y) - \beta I(y; z),$$

where x denotes the original text, y the mapped text, and z the downstream task variable.

Let the difference between the objectives of the semantic-decoupling mapping and the semantic-preserving mapping be

$$\Delta = \mathcal{G}(\mathcal{M}_{\mathrm{SD}}) - \mathcal{G}(\mathcal{M}_{\mathrm{SP}}).$$

Expanding the definition yields

$$\begin{aligned}\Delta = &\big(I^{\mathrm{SD}}(x;y) - I^{\mathrm{SP}}(x;y)\big) \\ &- \beta\big(I^{\mathrm{SD}}(y;z) - I^{\mathrm{SP}}(y;z)\big).\end{aligned}$$

Define the compression gain

$$\Delta C = I^{\mathrm{SP}}(x;y) - I^{\mathrm{SD}}(x;y),$$

and the task-information difference

$$\Delta T = I^{\mathrm{SP}}(y;z) - I^{\mathrm{SD}}(y;z).$$

Then

$$\Delta = -\Delta C + \beta \Delta T.$$

Therefore,

$$\mathcal{G}(\mathcal{M}_{\mathrm{SD}}) < \mathcal{G}(\mathcal{M}_{\mathrm{SP}})$$

holds whenever

$$\Delta C > \beta \Delta T.$$

Compression term. Assume that the semantic and structural components of the text are independent,

$$x = (x_s, x_c), \qquad x_s \perp x_c.$$

Under this assumption,

$$I(x;y) = I(x_s;y_s) + I(x_c;y_c).$$

Using the definitions of preservation ratios,

$$I(x;y) = \rho_s H(x_s) + \rho_c H(x_c).$$

The compression term is therefore

$$\begin{aligned}\Delta C = &(\rho_s^{SP} - \rho_s^{SD}) H(x_s) \\ &+ (\rho_c^{SP} - \rho_c^{SD}) H(x_c).\end{aligned}$$

Using the definitions

$$\Delta\rho_s = \rho_s^{SP} - \rho_s^{SD},\ \Delta\rho_c = \rho_c^{SP} - \rho_c^{SD},$$

this becomes

$$\Delta C = \Delta\rho_s H(x_s) + \Delta\rho_c H(x_c).$$

Under the independence assumption, there is

$$H(x) = H(x_s) + H(x_c).$$

Define the semantic information ratio

$$r = \frac{H(x_s)}{H(x)}.$$

Then

$$\frac{H(x_s)}{H(x)} = r, \qquad \frac{H(x_c)}{H(x)} = 1 - r.$$

Thus

$$\Delta C = H(x)\big(r\Delta\rho_s + (1-r)\Delta\rho_c\big).$$

Task-information term. The mutual information between the mapped text and the downstream task can be decomposed as

$$I(y; z) = I(y_s; z) + I(y_c; z).$$

From the definitions of the task-relevant preservation ratios,

$$I(y; z) = \eta_s I(x_s; z) + \eta_c I(x_c; z).$$

Using the task dependence coefficients

$$\lambda_s = \frac{I(x_s; z)}{I(x; z)}, \qquad \lambda_c = \frac{I(x_c; z)}{I(x; z)},$$

there is

$$I(y; z) = (\eta_s \lambda_s + \eta_c \lambda_c) I(x; z).$$

Therefore

$$\begin{aligned} \Delta T &= (\eta_s^{SP} \lambda_s + \eta_c^{SP} \lambda_c) I(x; z) \\ &\quad - (\eta_s^{SD} \lambda_s + \eta_c^{SD} \lambda_c) I(x; z) \\ &= (\Delta\eta_s \lambda_s + \Delta\eta_c \lambda_c) I(x; z), \end{aligned}$$

where

$$\Delta\eta_s = \eta_s^{SP} - \eta_s^{SD}, \qquad \Delta\eta_c = \eta_c^{SP} - \eta_c^{SD}.$$

Combining the two terms. The condition

$$\Delta C > \beta \Delta T,$$

becomes

$$\begin{aligned} &H(x)\big(r\Delta\rho_s + (1-r)\Delta\rho_c\big) \\ &> \beta(\Delta\eta_s \lambda_s + \Delta\eta_c \lambda_c) I(x; z). \end{aligned}$$

Let

$$\beta' = \beta \cdot \frac{I(x; z)}{H(x)}.$$

Dividing both sides by $H(x)$ gives

$$r\Delta\rho_s + (1-r)\Delta\rho_c > \beta'(\Delta\eta_s \lambda_s + \Delta\eta_c \lambda_c).$$

Since

$$\lambda_c = 1 - \lambda_s,$$

the right-hand side becomes

$$\beta'(\Delta\eta_s \lambda_s + \Delta\eta_c (1 - \lambda_s)).$$

Solving the inequality for λ_s yields

$$\lambda_s < \frac{r\Delta\rho_s + (1-r)\Delta\rho_c - \beta'\Delta\eta_c}{\beta'(\Delta\eta_s - \Delta\eta_c)}.$$

Whenever this condition holds, the inequality

$$\mathcal{G}(\mathcal{M}_{\mathrm{SD}}) < \mathcal{G}(\mathcal{M}_{\mathrm{SP}}),$$

is satisfied. This completes the proof.

C Analysis of Attack Success Rate

This section supplements the analysis of the relationship between KL divergence and attack success rate in Section 4.

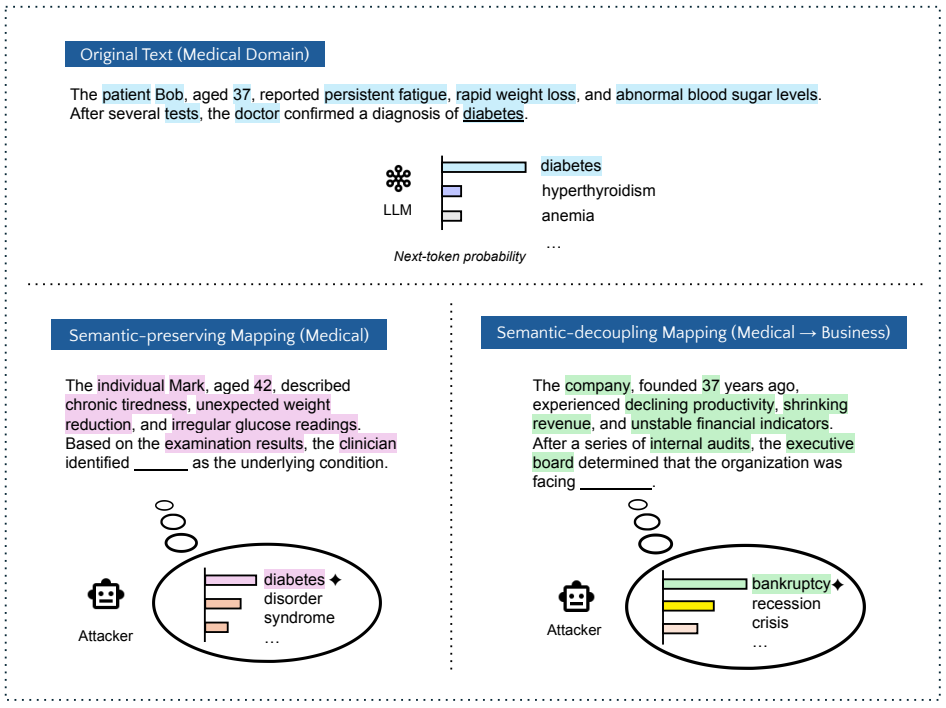


Figure 3: Examples of semantic-preserving and semantic-decoupling mappings under posterior sampling attacks. Semantic decoupling prevents correct inferences by stripping contextual cues from the source domain.

Posterior-based span inference attack. Let x be the original text, and let $s^\star := (s_1^\star, \ldots, s_m^\star)$ denote the m sensitive spans in x protected by the mapping mechanism. Given the attacker-observed output $y = \mathcal{M}(x)$, the attacker aims to recover each hidden span $s_i^\star$ from y.

Let $q_\theta(s_i|y)$ denote the attacker model, e.g., an LLM posterior over the i-th span conditioned on y. For each span $i \in \{1, \ldots, m\}$, the attacker outputs K guesses

$$\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)} \sim q_\theta(s_i|y).$$

Equivalently, the attacker forms a candidate set

$$\widehat{\mathcal{C}}_i^K(y) := \{\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)}\}.$$

A span is said to be successfully recovered if the ground-truth span appears in the candidate set, i.e.,

$$s_i^\star \in \widehat{\mathcal{C}}_i^K(y).$$

Attack success rate at top-K. We define the per-example attack success rate at top-K as the fraction of sensitive spans successfully recovered:

$$\mathrm{ASR@}K(x, y) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left\{ s_i^\star \in \widehat{\mathcal{C}}_i^K(y) \right\}.$$

Taking expectation over the data distribution and mapping randomness gives

$$\mathrm{ASR@}K(\mathcal{M}) := \mathbb{E}_{x \sim p(x),\, y \sim \mathcal{M}(x)} \left[\mathrm{ASR@}K(x, y) \right].$$

Thus, $\mathrm{ASR@}K(\mathcal{M})$ measures the expected proportion of sensitive spans that can be recovered by the attacker using K guesses per span.

Proposition 1 ($\mathrm{ASR@}K$ under posterior sampling)**.** *Assume that for each sensitive span i, the attacker draws K i.i.d. guesses from the posterior $q_\theta(s_i|y)$. Then*

$$\mathrm{ASR@}K(\mathcal{M}) = \mathbb{E}_{x,y}\left[\frac{1}{m}\sum_{i=1}^{m}\Big(1-\big(1-q_\theta(s_i^\star \mid y)\big)^K\Big)\right].$$

Proof. For a fixed (x, y) and a fixed span i, let

$$p_i(y) := q_\theta(s_i^\star \mid y).$$

Since the attacker draws K i.i.d. guesses from $q_\theta(s_i|y)$, the probability that none of the K guesses equals the true span $s_i^\star$ is

$$(1-p_i(y))^K.$$

Therefore, the probability that the true span is recovered at least once is

$$1-(1-p_i(y))^K = 1-\big(1-q_\theta(s_i^\star \mid y)\big)^K.$$

Taking the average over the m spans gives

$$\mathbb{E}[\mathrm{ASR@}K(x,y) \mid x,y] = \frac{1}{m}\sum_{i=1}^{m}\Big(1-\big(1-q_\theta(s_i^\star \mid y)\big)^K\Big).$$

Finally, taking expectation over (x, y) yields the result. □

Connection to KL divergence. For each sensitive span $s_i^\star$, define its surprisal (Kakouros et al., 2023) under the attacker posterior as

$$\mathcal{S}_i(y) := -\log_2 q_\theta(s_i^\star \mid y).$$

Taking expectation over the data distribution

$$\begin{aligned}\mathbb{E}\big[\mathcal{S}_i(\mathcal{Y})\big] &= H(\mathcal{S}_i(\mathcal{Y}) \mid \mathcal{Y})\\ &= H(\mathcal{S}_i(\mathcal{Y})) - I(\mathcal{S}_i(\mathcal{Y});\mathcal{Y}),\end{aligned}$$

where

$$I(\mathcal{S}_i(\mathcal{Y});\mathcal{Y}) = \mathbb{E}_y\big[\mathrm{KL}\big(p(s_i \mid y)\,\|\,p(s_i)\big)\big].$$

Therefore, smaller KL divergence implies larger expected surprisal. Furthermore, since

$$q_\theta(s_i^\star \mid y) = 2^{-\mathcal{S}_i(y)},$$

then Proposition 1 can be rewritten as

$$\mathrm{ASR@}K(\mathcal{M}) = \mathbb{E}_{x,y}\left[\frac{1}{m}\sum_{i=1}^{m}\Big(1-\big(1-2^{-\mathcal{S}_i(y)}\big)^K\Big)\right].$$

Therefore, $\mathrm{ASR@}K$ is a monotonically decreasing function of the span surprisal. Lower surprisal assigns larger posterior mass to the ground-truth span, which increases the probability that the true span appears among the attacker's K guesses. Consequently, a larger posterior-to-prior KL divergence leads to smaller expected surprisal and hence higher $\mathrm{ASR@}K$.

This aligns with the qualitative conclusion of this section: semantic-preserving mappings tend to be more vulnerable because they preserve cues in y that concentrate the posterior around $s^\star$. Figure 3 provides an example of such pattern.

D ALGORITHM

The details of the inference workflow with CROSS-MAP described in Section 5.3 are summarized in Algorithm 1.

Algorithm 1 Cross-domain Mapping and Recovery

Require: Trained mapper $\mathcal{M}_\phi$; trained recoverer $\mathcal{R}_\psi$; downstream $\mathcal{LLM}$.
1: **for** each input x **do**
2: $(\tau, D, y) \sim \mathcal{M}_\phi(x)$.
3: $y_{\mathrm{a}} \leftarrow \mathcal{LLM}(y)$.
4: $(\tilde{x}, \tilde{x}_{\mathrm{a}}) \sim \mathcal{R}_\psi\big(\tau, D, \mathrm{concat}(y, y_{\mathrm{a}})\big)$.
5: **return** $\tilde{x}_{\mathrm{a}}$.
6: **end for**

E SECURITY ANALYSIS OF CROSS-MAP

This section analyzes the security of CROSS-MAP against posterior-based attackers under dictionary leakage, when the attacker additionally observes or infers a subset of the mapper dictionary.

As discussed in Appendix C, the security of a mapping mechanism can be characterized through the posterior mass assigned to the ground-truth sensitive spans. In particular, under posterior-based attacks with K guesses per span, the attack success rate $\mathrm{ASR@}K$ is determined by the posterior probabilities of the true spans, equivalently by their surprisal. We now analyze how these quantities change when part of the dictionary is leaked to the attacker.

Leakage model. Let $s^\star = (s_1^\star, \ldots, s_m^\star)$ denote the m sensitive spans in the original text x, and let $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_m)$ be their mapped-domain counterparts produced by CROSS-MAP. The full mapper dictionary for this example is

$$D = \{(s_i^\star, \hat{s}_i)\}_{i=1}^m.$$

Assume the attacker additionally learns a leaked subset $D_{\mathrm{leak}} \subseteq D$ containing ℓ revealed pairs, with leakage rate $\rho = \ell/m$. Let $\mathcal{I}_{\mathrm{leak}} \subseteq \{1, \ldots, m\}$ denote the indices of leaked spans, and let $\mathcal{I}_{\mathrm{hid}} := \{1, \ldots, m\} \setminus \mathcal{I}_{\mathrm{leak}}$ be the indices of the still-hidden spans. Given the attacker-observed mapped output $y = \mathcal{M}(x)$, dictionary leakage enriches the attacker-visible information from y to (y, D_{leak}).

Accordingly, for each hidden span $i \in \mathcal{I}_{\mathrm{hid}}$, the attacker posterior changes from $q_\theta(s_i \mid y)$ to the leakage-conditioned posterior $q_\theta^\rho(s_i \mid y, D_{\mathrm{leak}})$. For leaked spans $i \in \mathcal{I}_{\mathrm{leak}}$, recovery is trivial once the corresponding dictionary pair is known.

Attack success rate under leakage. Under the same attack protocol as in Appendix C, the attacker outputs K guesses for each span. For leaked spans, the correct value is directly recovered from D_{leak}. For hidden spans, the attacker draws K guesses from the leakage-conditioned posterior $q_\theta^\rho(s_i|y, D_{\mathrm{leak}})$. Define the per-example success rate under leakage as

$$\mathrm{ASR@}K_\rho(x, y) := \frac{1}{m}\left(\sum_{i \in \mathcal{I}_{\mathrm{leak}}} 1 + \sum_{i \in \mathcal{I}_{\mathrm{hid}}} \mathbf{1}\Big\{s_i^\star \in \widehat{\mathcal{S}}_{i,\rho}^K(y)\Big\}\right),$$

where $\widehat{\mathcal{S}}_{i,\rho}^K(y)$ denotes the attacker's K guesses for span i under leakage. Taking expectation over the data distribution and mapping randomness gives

$$\mathrm{ASR@}K_\rho(\mathcal{M}) := \mathbb{E}_{x,y}[\mathrm{ASR@}K_\rho(x, y)]\,.$$

Proposition 2 ($\mathrm{ASR@}K_\rho$ under dictionary leakage)**.** *Assume that for each hidden span $i \in \mathcal{I}_{\mathrm{hid}}$, the attacker draws K i.i.d. guesses from $q_\theta^\rho(s_i|y, D_{\mathrm{leak}})$. Then*

$$\mathrm{ASR@}K_\rho(\mathcal{M}) = \mathbb{E}_{x,y}\Bigg[\frac{1}{m}\Big(|\mathcal{I}_{\mathrm{leak}}| + \sum_{i \in \mathcal{I}_{\mathrm{hid}}}\Big(1 - \big(1 - q_\theta^\rho(s_i^\star \mid y, D_{\mathrm{leak}})\big)^K\Big)\Big)\Bigg].$$

Proof. For each leaked span $i \in \mathcal{I}_{\text{leak}}$, the true span is directly revealed by the leaked dictionary pair, so its contribution to the success rate is 1.

For each hidden span $i \in \mathcal{I}_{\text{hid}}$, let

$$p_i^\rho(y) := q_\theta^\rho(s_i^\star \mid y, D_{\text{leak}}).$$

Since the attacker draws K i.i.d. guesses from the leakage-conditioned posterior, the probability that none of the K guesses equals the ground-truth span is

$$(1 - p_i^\rho(y))^K.$$

Therefore, the probability that the true span is recovered at least once is

$$1 - (1 - p_i^\rho(y))^K = 1 - \left(1 - q_\theta^\rho(s_i^\star \mid y, D_{\text{leak}})\right)^K.$$

Averaging over all m spans yields the conditional per-example success rate, and taking expectation over (x, y) gives the result. □

Connection to surprisal under leakage. For each hidden span $i \in \mathcal{I}_{\text{hid}}$, define its leakage-conditioned surprisal as

$$\mathcal{S}_i^\rho(y) := -\log_2 q_\theta^\rho(s_i^\star \mid y, D_{\text{leak}}).$$

Then Proposition 2 can be rewritten as

$$\text{ASR@}K_\rho(\mathcal{M}) = \mathbb{E}_{x,y}\left[\frac{1}{m}\left(|\mathcal{I}_{\text{leak}}| + \sum_{i \in \mathcal{I}_{\text{hid}}} \left(1 - \left(1 - 2^{-\mathcal{S}_i^\rho(y)}\right)^K\right)\right)\right].$$

Hence, for the unrevealed spans, dictionary leakage increases attack success exactly when it decreases the surprisal of the ground-truth spans under the attacker's posterior.

Effect 1: Search-space reduction for hidden spans. Dictionary leakage reduces the effective search space for the remaining hidden spans. Intuitively, once some span–mapping pairs are known, candidates inconsistent with the leaked dictionary can be eliminated, which increases the relative posterior mass assigned to the feasible values of the unrevealed spans.

To make this precise, suppose that for each hidden span i, the attacker originally considers a feasible candidate set $\mathcal{C}_i(y)$, while under leakage the feasible set shrinks to

$$\mathcal{C}_i^\rho(y, D_{\text{leak}}) \subseteq \mathcal{C}_i(y).$$

If the posterior over the feasible set is approximately uniform, then

$$q_\theta(s_i^\star \mid y) \approx \frac{1}{|\mathcal{C}_i(y)|},$$

$$q_\theta^\rho(s_i^\star \mid y, D_{\text{leak}}) \approx \frac{1}{|\mathcal{C}_i^\rho(y, D_{\text{leak}})|}.$$

Since

$$|\mathcal{C}_i^\rho(y, D_{\text{leak}})| \leq |\mathcal{C}_i(y)|,$$

the posterior mass of the true span increases after leakage, which lowers surprisal and raises $\text{ASR@}K$.

Effect 2: Posterior concentration via cross-span dependence. Beyond search-space reduction, leaked spans can provide additional semantic or structural clues about the remaining hidden spans. Let s_{hid} and s_{leak} denote the collections of hidden and leaked spans, respectively. Conditioning on the leaked dictionary cannot increase uncertainty:

$$H(s_{\text{hid}} \mid y, D_{\text{leak}}) \leq H(s_{\text{hid}} \mid y).$$

Equivalently,

$$\begin{aligned} &I(s_{\text{hid}}; D_{\text{leak}} \mid y) \\ &= H(s_{\text{hid}} \mid y) - H(s_{\text{hid}} \mid y, D_{\text{leak}}) \geq 0. \end{aligned}$$

Thus, leaked dictionary entries provide additional information about the unrevealed spans beyond what is already contained in y. At the per-span level, for each hidden span s_i this implies

$$H(s_i \mid y, D_{\text{leak}}) \leq H(s_i \mid y).$$

If the attacker posterior is well calibrated to the true posterior, then the expected leakage-conditioned surprisal satisfies

$$\begin{aligned} \mathbb{E}[\mathcal{S}_i^{\rho}(\mathcal{Y})] &= H(s_i \mid \mathcal{Y}, D_{\text{leak}}) \\ &\leq H(s_i \mid \mathcal{Y}) = \mathbb{E}[\mathcal{S}_i(\mathcal{Y})]. \end{aligned}$$

Therefore, dictionary leakage reduces the expected surprisal of the unrevealed spans and increases their recovery probability under posterior-based attacks.

The above analysis shows that dictionary leakage harms security through two coupled mechanisms: it directly reveals the leaked spans themselves, and it indirectly increases the recoverability of the remaining spans by concentrating the attacker posterior. In terms of the CROSS-MAP framework developed in this paper, both effects increase the posterior mass on the ground-truth spans, reduce their surprisal, and therefore raise $\text{ASR@}K$. This motivates evaluating security as a function of the leakage rate ρ and reporting robustness against strong attackers with partial dictionary knowledge inferred from historical observations.

F Supplementary Experimental Results

This section provides supplementary details on the experimental setup and results.

F.1 Experimental Setup

This subsection reports the experimental setup, including datasets, baselines, models, and evaluation metrics.

Datasets. SQuAD (Rajpurkar et al., 2016), BioASQ (Tsatsaronis et al., 2015), FinCausal 2025 (Moreno-Sandoval et al., 2025), and NarrativeQA (Kočiskỳ et al., 2018) are used for QA-style evaluation across general, biomedical, financial, and narrative domains. CNN/DailyMail is used for summarization, and CommonGen is used for constrained open-ended generation. SynthPAI (Yukhymenko et al., 2024) is used for evaluation against de-anonymization attacks. Model optimization involves a subset of SQuAD comprising 754 instances for SFT and 1,502 instances for DPO. For QA evaluation, 500 test instances are sampled from each dataset. The additional summarization and open-ended generation experiments use 200 CNN/DailyMail instances and 100 CommonGen instances, respectively. All reported results are averaged across the corresponding test subset in a single run for each dataset.

Baselines. State-of-the-art methods for inference-time privacy protection are used as baselines, including the sanitization method InferDPT (Tong et al., 2025b), and bidirectional frameworks HAS (Chen et al., 2023b) and Prεεmpt (Chowdhury et al., 2025). For evaluation against de-anonymization attacks, the adversarial anonymization framework GPT-AA (Staab et al., 2025) is also included as a baseline.

Metrics. The evaluation focuses on the trade-off between utility and privacy across tasks. For QA, utility is assessed with EM/F1/ACC. For summarization, utility is assessed with ROUGE-L and BERTScore. For CommonGen, utility is assessed with ROUGE-L and concept coverage. The utility-oriented training objectives introduced in Section 5, including fluency and recovery quality, are also reported where applicable. Privacy is measured using dictionary-/text-level distance, dictionary coverage, sensitive-span retention, and the posterior-mass quantity derived from surprisal in Appendix E. Beyond these intrinsic metrics, attack success rate (ASR) is reported under different attackers to reflect effective privacy under each threat model.

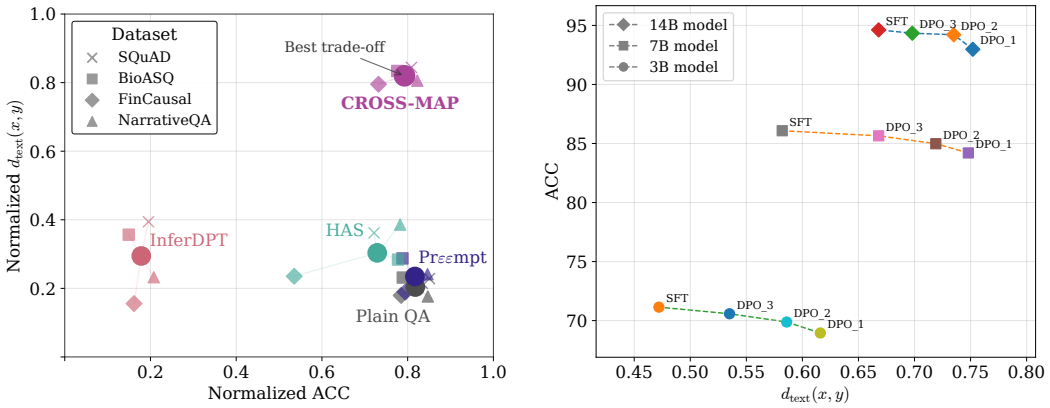
Models. Qwen-2.5-3B/7B/14B (Qwen et al., 2025) and Qwen3-4B (Yang et al., 2025a) are used as the base models for the mapper and the recoverer. LLAMA-8B is used as the downstream model for the main experiments, and the additional deployment ablation also evaluates stronger external LLM access through CROSS-MAP. LoRA is used for parameter-efficient fine-tuning with $r = 16$, $\alpha = 32$, and a learning rate of 1×10^{-5} for all runs.

Prompts. Figure 7 and 8 show the mapper and recoverer prompts, respectively.

F.2 Supplementary Experimental Results

This section provides a full description of experimental results, including tables and figures, as a supplement to Section 6.

F.2.1 Utility and Privacy



(a) Privacy–utility trade-off for different methods.

(b) Privacy–utility trade-off for different training stages.

Figure 4: Privacy–utility trade-offs. In (b), dashed lines connect checkpoints within each model size. DPO_3, DPO_2, and DPO_1 denote checkpoints from epochs 1, 3, and 5.

Table 3: Generalization on summarization and generation tasks. The BERT/Cov. metric denotes BERTScore for summarization and concept coverage for generation. Retention represents the sensitive-span retention rate.

Task	Method	ROUGE-L	BERT/Cov.	Retention
CNN/DM	Plain LLM	0.240	0.882	100.0%
CNN/DM	CROSS-MAP	0.285	0.877	7.3%
CommonGen	Plain LLM	0.587	0.932	98.3%
CommonGen	CROSS-MAP	0.414	0.891	18.9%

In Table 2 of the main paper, the ACC metric calculates the proportion of recovered LLM responses that contain the ground-truth answer span. The similarity (Sim.) metric for recovered text is defined as $1 - d_{\text{text}}(x, \tilde{x})$. The BLEU and ROUGE-1 (R-1) metrics are also computed by comparing x with

$\tilde{x}$. Fluency Loss (Flu. Loss) is the token-level negative log-likelihood under the reference LM, i.e., the negative of the fluency score f_{text} in Section 5.1. Other metrics are consistent with the definitions in Section 5.1.

Among all methods, Plain QA and Pr$\varepsilon\varepsilon$mpt retain the highest downstream utility but provide negligible or weak privacy. For Pr$\varepsilon\varepsilon$mpt, the mapped-text privacy scores remain low overall, reflecting its minimal perturbation on the original text. InferDPT delivers only partial privacy gains while causing a substantial utility collapse (e.g., on SQuAD, F1 drops from 84.10 to 74.04), highlighting that perturbation without an explicit recovery mechanism distorts answer-critical semantics. As a bidirectional framework, HAS attains higher utility than InferDPT, but its privacy remains below CROSS-MAP, suggesting that residual semantic cues are still preserved in the mapped text and thus remain exposable to external LLMs. Across all datasets, CROSS-MAP consistently achieves the strongest mapped-text privacy while largely preserving downstream QA utility relative to other baselines. For example, on SQuAD, the distance between mapped text and original text increases from HAS's 0.238 to 0.752 while QA accuracy increases from HAS's 91.84 to 92.98. Similar trends are observed on other datasets. These consistent gains indicate that high downstream utility does not require high similarity between mapped and original text. In summary, the consistent privacy gains and the near-preserved downstream utility achieved by CROSS-MAP demonstrate that a bidirectional framework with semantic decoupling yields the best privacy-utility trade-off among the compared methods.

Figure 4a further visualizes each method in terms of downstream QA accuracy and the text distance $d_{\text{text}}(x, y)$. Across datasets, Pr$\varepsilon\varepsilon$mpt clusters near Plain QA in the high-utility but low-privacy region, because it primarily targets numerical values and leaves most of the original text unchanged. InferDPT typically shifts upward in privacy but drifts leftward in utility due to its lack of an explicit recovery process. HAS provides a better trade-off than InferDPT and Pr$\varepsilon\varepsilon$mpt. However, it is still inferior to CROSS-MAP, which consistently lies closest to the upper-right corner, achieving larger $d_{\text{text}}(x, y)$ than other baselines, and meanwhile maintains downstream QA performance very close to Plain QA.

Table 4 reports LLM-as-a-judge ratings across different methods. The judge rates all metrics on a 1–5 scale, where higher scores in the first two columns indicate better plausibility (Plaus.) and fluency (Flu.), whereas lower scores in the last column indicate fewer conspicuous obfuscation artifacts for detectability (Detect.).

Table 4: LLM-as-a-judge ratings across methods.

Method	Plaus. ↑	Fluency ↑	Detect. ↓
Original text	4.93	4.91	1.06
Pr$\varepsilon\varepsilon$mpt	4.82	4.86	1.18
HAS	4.54	4.61	1.58
InferDPT	3.71	3.94	2.47
CROSS-MAP	4.45	4.52	1.72

F.2.2 De-anonymization Attacks

Compared to Plain Text, Pr$\varepsilon\varepsilon$mpt achieves only a marginal reduction in attack success (-1.7%) and confidence (-6.4%), indicating that limited edits remain insufficient to remove attribute cues. In contrast, HAS reduces inference accuracy by 34.1%, and GPT-AA further reduces it by 40.7%, accompanied by lowered certainty. The strongest defense is achieved by CROSS-MAP, where inference accuracy drops significantly by 88.5%. This pattern suggests that the semantic decoupling in CROSS-MAP suppresses attribute cues more thoroughly, leading to both lower attacker accuracy and lower confidence.

Table 5: De-anonymization results via LLM attribute inference on SynthPAI (Yukhymenko et al., 2024), reporting attacker inference accuracy and confidence (0-5).

Method	Inference Accuracy	Inference Certainty
Plain Text	0.713	3.366
Prεεmpt	0.701	3.152
HAS	0.470	2.196
GPT-AA	0.423	2.073
CROSS-MAP	**0.082**	**1.063**

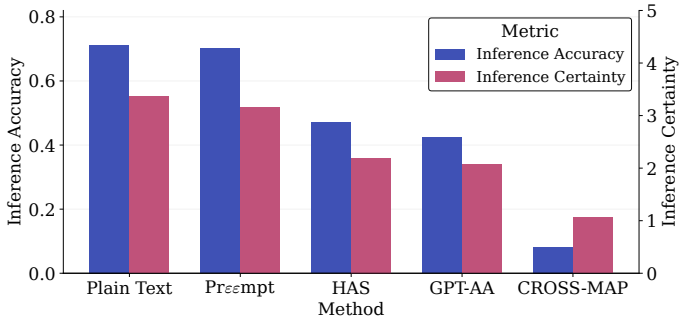


Figure 5: De-anonymization results via LLM attribute inference on SynthPAI (Yukhymenko et al., 2024), comparing attacker inference accuracy and certainty across methods.

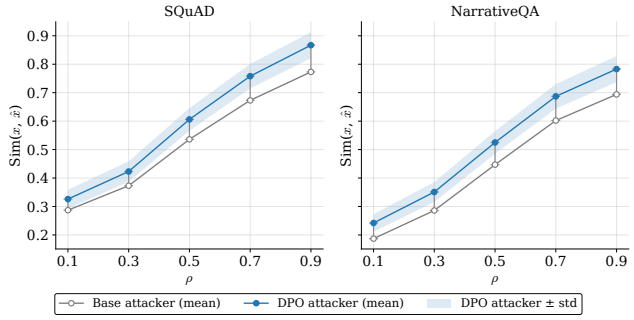


Figure 6: Optimization-based attacker's restoration rate under different dictionary leakage ratios ρ.

F.2.3 Span Restoration Attacks

Figure 2 provides direct empirical evidence that posterior sampling succeeds when the protected span remains a high-probability continuation under the attacker's token-level posterior. The privacy advantage of CROSS-MAP is therefore attributed to its semantic-decoupling mapping, where direct lexical and semantic anchoring between mapped text and the original protected spans is systematically weakened. As a result, the attacker's posterior over the protected spans is flattened. The true span no longer concentrates probability mass among top-k candidates, yielding lower posterior mass, equivalently higher surprisal, and lower ASR consistently across datasets.

F.2.4 Optimization-based Attacks

Table 13 and Table 14 show that restoration becomes easier as ρ increases for both datasets, and that the DPO-based attacker consistently outperforms the base attacker without DPO optimization at every ρ. For example, on SQuAD, the DPO-based attacker's mean similarity score increases from 0.326 to 0.867 as ρ goes from 0.1 to 0.9, with absolute mean gains over the base attacker ranging from 0.039 to 0.094. This systematic improvement suggests that optimization reliably pushes generations toward

better reconstruction, rather than merely benefiting from occasional lucky samples. Qwen2.5-3B is used for both attackers in these two tables.

From Figure 6, it can be clearly observed that the restoration similarity rises monotonically with ρ. Furthermore, higher leakage is also accompanied by larger absolute gains. A possible reason is that higher ρ reveals more dictionary information, sharpening the attacker's posterior, while DPO further concentrates probability mass on outputs that yield higher restoration similarity.

F.2.5 Ablation Studies

This section details an ablation analysis of key model and deployment factors from multiple perspectives: input complexity, paraphrase robustness, target-domain selection, dictionary coverage, training scale, external LLM capabilities, mapper/recoverer size, and training strategies.

Input complexity. Three parse-based metrics are used to measure syntactic complexity: mean dependency distance (MDD), maximum dependency-tree depth (max depth), and clauses per sentence (clauses/sent.). NarrativeQA has an average MDD, max depth, and clauses per sentence of 2.40, 12.8, and 2.95, respectively. SQuAD has $2.40/10.6/2.26$, BioASQ has $2.41/7.6/2.03$, and FinCausal has $1.98/7.7/1.93$.

Compared to the example sentence, *"Although the defendant had no prior convictions, the judge imposed a harsher sentence because the minor victim suffered irreversible harm"* (MDD $= 3.00$, max depth $= 6$, and clauses/sent. $= 4.00$), the datasets used for evaluation in this paper include multi-clause and long-distance syntactic structures.

Table 6: SQuAD complexity-quartile robustness. Examples are sorted by MDD and split into Q1–Q4, where Q4 is most complex.

Split	MDD	Max depth	Clauses/sent.	ACC ↑	Flu. Loss ↓
Q1	2.16	11.11	2.27	93.62%	4.722
Q2	2.40	10.75	2.34	90.52%	4.687
Q3	2.65	9.85	2.53	93.36%	4.715
Q4	3.24	10.11	2.90	94.42%	4.688

Table 6 reports the evaluation of complexity on the SQuAD dataset. The results show that increasing syntactic complexity does not monotonically degrade utility or fluency. Q4, the most complex quartile, reaches 94.42% ACC and a Fluency Loss of 4.688. This suggests that CROSS-MAP is robust to complex structures.

Paraphrasing Robustness. Table 7 evaluates CROSS-MAP's recovery robustness against the paraphrasing of mapped text. The results show that the ACC for SQuAD drops mildly from 0.929 to 0.877 under WordNet synonym substitution ($p = 0.30$) and to 0.896 under T5 paraphrasing. Such robustness to paraphrasing is important for third-party LLM APIs, where downstream generation cannot be forced to copy the mapper's terms exactly.

Table 7: Paraphrase robustness of the recoverer on SQuAD. The recoverer reasons over the dictionary and response jointly, rather than relying on exact string matching.

Setting	ACC	Drop
Original output	0.929	–
WordNet synonyms	0.877	5.2%
T5 paraphrase	0.896	3.3%

Domain selection. To investigate the interplay between semantic domain distance, privacy preservation, and task utility, we conduct an ablation analysis in Table 8. The within-cluster row averages d_{text} and ACC over samples mapped between source-target domain pairs in the same cluster. The across-cluster row averages over samples mapped between pairs from different clusters. The results

indicate that across-cluster mappings significantly augment d_{text} from 0.642 to 0.797, confirming that increased domain divergence strengthens semantic separation. Notably, this substantial privacy gain is achieved with only a minor reduction in utility (a 3.5% decrease in accuracy, shifting from 90.4% to 86.9%). This desirable trade-off suggests that the mapping can be executed with high quality despite a large semantic distance, provided that there is sufficient structural and relational alignment between the source and target domains to preserve the underlying discourse mechanics.

Table 8: Target-domain distance ablation over 260 source-target pairs. Source and target domains are embedded with Sentence-BERT and grouped into semantic-domain clusters.

Pairing	d_{text}	ACC
Within-cluster	0.642	90.4%
Across-cluster	0.797	86.9%

Table 9 further provides a qualitative case study for domain selection. The same cognitive-development source text is mapped to a far equipment-operations domain and a nearer medical-training domain. Both mappings preserve the key discourse functions: a long-range temporal frame, causal connective, contrastive predication, multi-clause explanation, and comparative grounding. However, the far-domain mapping needs more local reordering to remain natural, while the near-domain mapping can preserve more surface structure. This explains the quantitative pattern in Table 8: larger semantic distance can improve privacy, but the target domain needs enough compatible relational roles to support a fluent mapped passage and ensure accurate recovery.

Table 9: Domain-selection case study. Bold spans mark key source entities and their mapped counterparts.

Mapping	Text
Original: cognitive development	**Wisdom**, or the capacity for **insight and judgment** that is developed through **experience**, increases between the **ages of fourteen and twenty-five**, then levels off. Thus, it is during the **adolescence-adulthood transition** that individuals acquire the type of **wisdom** typically associated with age. **Wisdom** is not the same as **intelligence**: adolescents do not improve substantially on **IQ tests** because their scores are relative to others in their **same age group**, and relative standing usually does not change — everyone matures at approximately the same rate in this way.
Far domain: equipment operations	**Diagnostic maturity**, or the ability to **interpret sensor patterns and make maintenance decisions** refined through **operating hours**, improves between **300 and 1,800 operating hours**, then plateaus. Thus, units tend to develop the level of **diagnostic maturity** typically attributed to long service life during the transition from **commissioning to routine operation**. **Diagnostic maturity** is not the same as **rated capacity**: newly commissioned units do not show large gains on **standardized bench tests** because their results are normalized against **comparable models from the same production batch**, and relative ranking typically remains stable — most units stabilize at roughly similar rates under similar duty cycles on that metric.
Near domain: medical training	**Professional judgment**, or the capacity for **clinical insight and decision-making** that is developed through **supervised practice**, increases between the **first and final years of training**, then levels off. Thus, it is during the **trainee-practitioner transition** that clinicians acquire the type of **professional judgment** typically associated with seniority. **Professional judgment** is not the same as **medical knowledge**: trainees do not improve substantially on **standardized exams** because their scores are relative to others in their **same training cohort**, and relative standing usually does not change — everyone advances at approximately the same rate in this way.

Dictionary coverage. Table 10 ablates recovery behavior as dictionary coverage decreases. Recovery similarity decreases from 0.860 at approximately 0.8 coverage to 0.816 at approximately 0.5 coverage, and remains 0.788 at approximately 0.2 coverage. The trend is smooth rather than catastrophic because the recoverer can use the mapped response and task context to infer some missing links. However, the ACC drop from 91.2% to 86.2% also shows that the dictionary is not optional. It is the mechanism that keeps semantic decoupling recoverable for downstream use.

Table 10: Dictionary-coverage ablation on SQuAD.

Dictionary coverage	Recovery Sim.	ACC
$c_{\text{dict}} \approx 0.8$	0.860	91.2%
$c_{\text{dict}} \approx 0.5$	0.816	88.7%
$c_{\text{dict}} \approx 0.2$	0.788	86.2%

Training scale. Table 11 ablates SFT data size for the 14B model. Moving from the base model to 200 and 500 SFT examples already improves both privacy distance and utility, indicating that the desired behavior can be learned from a relatively small synthetic set. The largest utility gain appears by 700 examples, where ACC reaches 94.59 and recovery similarity reaches 0.945. Additional data from 1000 to 2000 examples yields only marginal changes in d_{text} and slightly lower recovery similarity, suggesting that later gains are dominated by DPO and sampling strategy rather than by simply scaling SFT data.

Table 11: SFT training-data scaling study. Gains largely saturate around 700–1000 samples.

Checkpoint	# samples	d_{text}	Sim. / ACC
Base	0	0.541	0.901 / 88.96
SFT-200	200	0.634	0.912 / 90.38
SFT-500	500	0.651	0.933 / 92.57
SFT-700	700	0.663	0.945 / 94.59
SFT-1000	1000	0.667	0.943 / 94.34
SFT-1500	1500	0.670	0.940 / 94.35
SFT-2000	2000	0.673	0.937 / 94.51

External LLMs. Table 12 compares local-only inference, unprotected external-LLM access, and external-LLM access through CROSS-MAP. The SQuAD rows use a 3B local mapper/recoverer and a Qwen-2.5-14B model as the external LLM, showing that CROSS-MAP remains useful even when the external model is a local open-weight model rather than a commercial API. The remaining rows use GPT-5 as the external LLM. In both regimes, CROSS-MAP retains most of the external model's utility while keeping original-domain sensitive content local.

Model Size and Training Strategy. Table 15 ablates training strategy (SFT-only vs. DPO) and mapper / recoverer size. Different DPO variants provide adjustable privacy-utility trade-offs for each model size. For the 14B model, as the DPO variant moves from $\text{DPO}_3 \rightarrow \text{DPO}_2 \rightarrow \text{DPO}_1$, privacy distances increase substantially, at the cost of slightly reduced QA utility. The strongest privacy setting DPO_1 (14B) raises d_{text} to 0.752 while keeping ACC at 92.98, i.e., only a -1.85% drop from Plain QA (94.73).

The SFT-only model provides weaker privacy than DPO_1 and a closer approximation of the plain QA quality. However, SFT-only exhibits substantially worse JSON-structure validity r_{json} than DPO variants, indicating that preference optimization also stabilizes format consistency, which is critical for reliable mapping and recovery. Figure 4b visualizes the same model-size and training-stage trade-off, with DPO checkpoints moving toward stronger privacy at small utility cost.

Table 12: External-LLM ablation. CROSS-MAP preserves most of the stronger external model's utility while keeping original-domain sensitive content local.

Setting	Dataset	Model / pipeline	EM	F1
Local only	SQuAD	Qwen-2.5-3B local	33.82	42.77
External, no protection	SQuAD	Qwen-2.5-14B external	54.28	62.51
CROSS-MAP	SQuAD	3B mapper/recoverer + 14B external	47.71	55.96
Local only	BioASQ	Qwen-2.5-14B local	12.63	23.35
External, no protection	BioASQ	GPT-5 external	24.98	46.43
CROSS-MAP	BioASQ	14B mapper/recoverer + GPT-5 external	22.36	44.57
Local only	FinCausal	Qwen-2.5-14B local	48.69	60.99
External, no protection	FinCausal	GPT-5 external	82.32	91.74
CROSS-MAP	FinCausal	14B mapper/recoverer + GPT-5 external	81.57	90.26
Local only	NarrativeQA	Qwen-2.5-14B local	16.72	39.91
External, no protection	NarrativeQA	GPT-5 external	48.75	72.94
CROSS-MAP	NarrativeQA	14B mapper/recoverer + GPT-5 external	47.12	71.27

Mapper Prompt

```
Task: Privacy-Preserving Mapping.
Input:
Original Prompt: {original_prompt}

Instructions:
1. Choose a non-sensitive target domain and provide a professional reason for your choice.
2. Create a 1-to-1 mapping dictionary covering key entities, actions, numbers, and concepts from the original
prompt.
3. Rewrite the original prompt into the chosen target domain.

Output Format: JSON object containing:
- "theme_selection_logic": "Reason for choosing the theme.",
- "theme": "The target domain name.",
- "dictionary": {{ "orig_term": "mapped_term", ... }},
- "mapped_prompt": "The full mapped prompt..."

CRITICAL REQUIREMENT:
1. For privacy, do NOT use simple synonym replacement. Use cross-domain transformation.
2. For restoration, all replacements must be kept in the dictionary.
3. Any numbers MUST be mapped to contextually reasonable different numbers and the mapping must be kept
in the dictionary.
4. Return ONLY valid JSON.
```

Figure 7: Mapper prompt.

Recoverer Prompt

```
Task: Privacy-Preserving Restoration.
Input:
Mapped Prompt: {mapped_prompt}
Mapped Response: {mapped response}
Dictionary: {dictionary}

Instructions:
1. Using the provided Dictionary, restore the Mapped Prompt and Mapped Response back to
their original domain.
2. The goal is to recover the original meaning and professional tone.

Output Format: JSON object containing:
- "restored_prompt": "The restored prompt in the original domain...",
- “restored_response”: ”The restored response in the original domain..."

Return ONLY valid JSON.
```

Figure 8: Recoverer prompt.

Table 13: Performance of optimization-based white-box attacker on SQuAD, measuring similarity between reconstructed text and original text across leakage ratios ρ.

ρ	Base attacker (mean)	DPO attacker (mean)	Gain (mean)	DPO attacker (best)	Std. (mean)
0.1	0.287	0.326	0.039	0.397	0.033
0.3	0.373	0.423	0.050	0.494	0.036
0.5	0.536	0.606	0.070	0.679	0.040
0.7	0.673	0.758	0.085	0.831	0.043
0.9	0.773	0.867	0.094	0.935	0.046

Table 14: Performance of optimization-based white-box attack on NarrativeQA, measuring similarity between reconstructed text and original text across leakage ratios ρ.

ρ	Base attacker (mean)	DPO attacker (mean)	Gain (mean)	DPO attacker (best)	Std. (mean)
0.1	0.187	0.242	0.055	0.295	0.039
0.3	0.286	0.351	0.065	0.418	0.045
0.5	0.447	0.525	0.078	0.607	0.049
0.7	0.602	0.687	0.085	0.772	0.051
0.9	0.694	0.783	0.089	0.832	0.046

Table 15: Ablation study on SQuAD, comparing DPO-tuned and SFT-only models of different sizes in terms of QA utility, privacy distances, and JSON-structure validity. LLAMA-8B model is used for downstream QA.

	Downstream QA utility			**Mapping quality**			
Method	EM	F1	ACC	Sim.	d_{dict}	d_{text}	r_{json}
Plain QA	76.4	84.10	94.73	1.000	0.000	0.000	1.000
DPO_1 (14B)	**74.94**	**83.22**	**92.98**	**0.826**	**0.753**	**0.752**	**1.000**
DPO_2 (14B)	75.30	83.31	94.21	0.892	0.711	0.735	1.000
DPO_3 (14B)	75.72	83.75	94.34	0.898	0.704	0.698	1.000
SFT-only (14B)	**75.88**	**83.82**	**94.62**	**0.948**	**0.612**	**0.668**	**0.936**
Base Model (14B)	69.85	78.90	88.96	0.901	0.566	0.541	0.988
DPO_1 (7B)	**70.21**	**81.88**	**84.20**	**0.829**	**0.758**	**0.748**	**1.000**
DPO_2 (7B)	70.74	82.31	84.98	0.844	0.736	0.719	1.000
DPO_3 (7B)	71.86	83.05	85.65	0.861	0.708	0.668	1.000
SFT-only (7B)	**71.20**	**82.74**	**86.08**	**0.901**	**0.604**	**0.582**	**0.918**
Base Model (7B)	62.14	76.82	73.44	0.872	0.522	0.446	0.972
DPO_1 (3B)	**54.21**	**69.35**	**68.95**	**0.775**	**0.624**	**0.616**	**1.000**
DPO_2 (3B)	54.88	69.87	69.88	0.788	0.603	0.586	1.000
DPO_3 (3B)	56.42	70.96	70.58	0.803	0.575	0.535	1.000
SFT-only (3B)	**55.69**	**70.34**	**71.14**	**0.846**	**0.401**	**0.472**	**0.782**
Base Model (3B)	46.10	64.88	60.73	0.781	0.365	0.322	0.958

Table 16: Examples of different mapping methods.

Method	Original text	Mapped text	Recovered text
CROSS-MAP	Fiscal pressures on Government as a result of reduced revenues due to lower oil prices is evident in the increasing number of audits being undertaken to identify areas of non-compliance. The Group remains committed to maintaining the highest levels of compliance and works closely with local regulatory authorities.	Operational strain on the hospital network as a result of declining patient volume due to reduced insurance reimbursements is evident in the increasing number of quality assurance reviews being undertaken to identify deviations from clinical protocols. The hospital network remains committed to adhering strictly to regulatory standards and works closely with health department oversight bodies.	Fiscal pressures on the Government as a result of reduced revenues due to lower oil prices is evident in the increasing number of audits being undertaken to identify non-compliance with regulatory standards. The Government remains committed to maintaining the highest levels of compliance and works closely with local regulatory authorities.
HAS (Chen et al., 2023b)	Fiscal pressures on Government as a result of reduced revenues due to lower oil prices is evident in the increasing number of audits being undertaken to identify areas of non-compliance. The Group remains committed to maintaining the highest levels of compliance and works closely with local regulatory authorities.	Fiscal pressures on Administration as a result of reduced revenues due to lower oil prices is evident in the increasing number of audits being undertaken to identify areas of non-compliance. The Committee remains committed to maintaining the highest levels of compliance and works closely with local regulatory authorities.	Fiscal pressures on Government as a result of reduced revenues due to lower oil prices is evident in the increasing number of audits being undertaken to identify areas of non-compliance. The Committee remains committed to maintaining the highest levels of compliance and works closely with local regulatory authorities.
InferDPT (Tong et al., 2025b)	In 2018 the majority of employees across the Group have received average salary increases ranging from 2.0%–3.25%, dependent on geographical location with the principal exception being those employees based in Brazil, Latin America and China where, due to inflation, current market salary increases are higher.	In 2019 the majority of employees across the Company have received average salary rises ranging from 2.0%–3.25%, dependent on geographical location with the principal exception being those employees based in Brazil, Latin America and China where, due to inflation, current market salary rises are higher.	In 2019 the majority of employees across the Company have received average salary rises ranging from 2.0%–3.25%, dependent on geographical location with the principal exception being those employees based in Brazil, Latin America and China where, due to inflation, current market salary rises are higher.
Prεεmpt (Chowdhury et al., 2025)	The Committee also reviewed the impact of the reduction in US federal tax rates as a result of tax reform in the US, which resulted in a reduction of deferred tax balances of 617 million.	The Committee also reviewed the impact of the reduction in US federal tax rates as a result of tax reform in the US, which resulted in a reduction of deferred tax balances of 593 million.	The Committee also reviewed the impact of the reduction in US federal tax rates as a result of tax reform in the US, which resulted in a reduction of deferred tax balances of 617 million.

Table 17: Black-box span restoration results across datasets, reporting span restoration accuracy as ASR and the attacker's posterior mass on the ground-truth span at different candidate list sizes k.

Dataset	Method	$k = 10$		$k = 30$		$k = 50$	
		ASR	Post. Mass	ASR	Post. Mass	ASR	Post. Mass
FinCausal 2025 (Moreno-Sandoval et al., 2025)	**CROSS-MAP**	**0.37**%	**0.006**	**0.84**%	**0.003**	**1.19**%	**0.002**
	HAS	5.88%	0.116	35.29%	0.026	35.29%	0.026
	Plain QA	7.86%	0.181	41.43%	0.088	41.43%	0.088
BioASQ (Tsatsaronis et al., 2015)	**CROSS-MAP**	**1.05**%	**0.004**	**1.05**%	**0.004**	**1.05**%	**0.004**
	HAS	12.73%	0.279	28.14%	0.164	34.66%	0.135
	Plain QA	16.29%	0.316	32.81%	0.239	36.12%	0.208
SQuAD (Rajpurkar et al., 2016)	**CROSS-MAP**	**0.68**%	**0.068**	**1.22**%	**0.042**	**1.49**%	**0.023**
	HAS	4.95%	0.164	12.10%	0.145	14.57%	0.124
	Plain QA	5.33%	0.209	15.70%	0.171	15.70%	0.171