TASK: TASK-AWARE, SECURITY-ITERATIVE PROMPT SANITIZATION FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Cloud-based large language models enable users to complete complex document-processing tasks without local deployment. However, prompts must leave the local environment during transmission to the cloud, creating substantial privacy risks. Existing prompt-sanitization methods commonly address this problem through substitution, obfuscation, or LLM-based rewriting. Such methods may excessively replace sensitive content or rely on rewriting criteria that do not align with the requirements of the current task, resulting in the loss of task-required information and substantially degrading model output quality. To address this problem, we make two key observations: spans of the same sensitive type can play different semantic roles across tasks, and residual risk after initial sanitization is typically concentrated in a small number of retained or weakly transformed spans and their associated evidence. Building on these observations, we propose TASK, a task-aware iterative prompt sanitization framework for cloud-based LLM inference. TASK jointly analyzes the semantic relevance of sensitive spans to the task and their local context to identify information that genuinely supports task completion. It then selects multi-level transformations, including retention, abstraction, generalization, and masking. To further reduce residual risk, TASK evaluates initially sanitized prompts using sensitive-value recovery and attribute-inference attacks, and iteratively adjusts high-risk decisions until the result satisfies the required privacy level. We evaluate TASK on document summarization, narrative question answering, and long-context question answering. The results show that TASK improves task quality and privacy scores by 23.72% and 8.64% on average, respectively.

1 INTRODUCTION

Large language models (LLMs) (OpenAI et al., 2024; Guo et al., 2025) are increasingly used to process documents in privacy-sensitive application scenarios. When users submit these documents to cloud-based LLM services, the data leave the local environment and may be exposed to third-party infrastructure, creating substantial privacy risks (Singhal et al., 2023; Wu et al., 2023). A common protection strategy is to sanitize sensitive information before transmission so that cloud-based LLMs cannot directly access the original sensitive content (Hou et al., 2025). However, directly removing sensitive information can damage the structure and semantics of the prompt.

To address this challenge, researchers have proposed prompt redaction (Aahill, 2023; Kan et al., 2023), which replaces original sensitive terms with generalized representations while preserving the basic syntactic structure. However, fixed substitution rules leave contextual cues from which private information can still be inferred. Another line of work, prompt obfuscation (Pape et al., 2025; Lin et al., 2025a), transforms the token-level or embedding-level representation of a prompt to reduce direct exposure of the original text. Its effectiveness, however, depends heavily on the perturbation configuration, and excessive perturbation can degrade model output quality. More recently, LLM Context Aware methods (Shen et al., 2026; Siyan et al., 2025; Yang et al., 2025; Pasch & Cha, 2025; Miranda et al., 2025) have employed language models to rewrite prompts based on contextual semantics, allowing sensitive content to be processed selectively while reducing semantic loss.

Although LLM Context Aware methods preserve more semantic information than prompt substitution and obfuscation, they still have limitations in terms of output quality and privacy protection.

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.

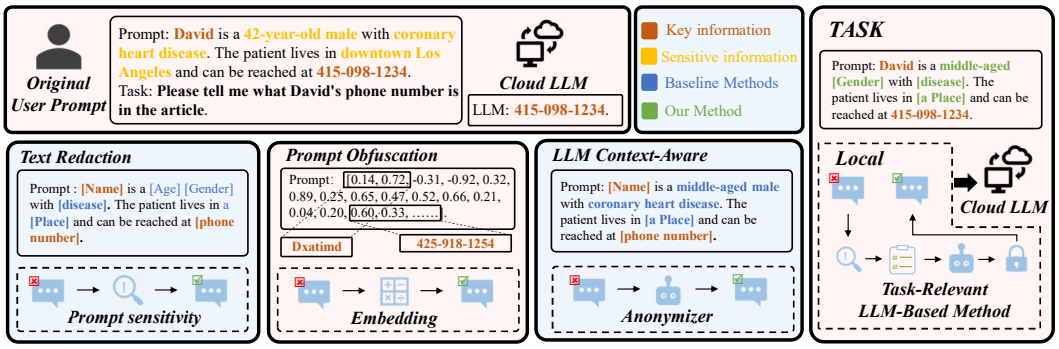


Figure 1: Comparison between existing methods and TASK. Prompt Redaction directly removes sensitive spans, which readily leads to information loss. Prompt Obfuscation alters the input representation, reducing the model's ability to understand the prompt. LLM Context Aware methods rewrite sensitive content based on contextual intent but fail to adequately reflect task requirements. In contrast, TASK jointly considers the task relevance and privacy risk of sensitive spans and iteratively refines sanitization decisions, thereby reducing privacy leakage while preserving task-relevant information.

As shown in Figure 1, existing methods (Frikha et al., 2025; Staab et al., 2024; Gardiner et al., 2024; Żarski & Janicki, 2025) generally select content for transformation according to its context or communicative intent, without explicitly assessing the semantic role of sensitive information in the current task. Moreover, they do not evaluate the sanitized prompt for residual disclosure risk. An adversary may therefore recover sensitive values or infer private attributes from the remaining contextual cues. Preserving task-relevant semantics while mitigating residual privacy risks consequently remains a key challenge.

To address this bottleneck, we analyze the relationships among sensitive spans, task semantics, and residual disclosure risks. This analysis yields two key observations. First, *even spans belonging to the same sensitive category can play substantially different roles in the current task.* Task-relevant spans are often strongly connected to their surrounding context, and removing them can cause substantial information loss. Second, *residual risk after initial sanitization is not uniformly distributed across all spans*; instead, it is concentrated in a small number of retained or weakly transformed spans and their combinations of supporting evidence.

Building on these observations, we propose a task-aware iterative prompt sanitization framework that combines task-relevance assessment of sensitive spans with iterative security evaluation of the sanitized prompt, as shown in Figure 1. The framework first identifies candidate sensitive spans and extracts their contextual and associated evidence. It then evaluates the semantic relevance of each span to the current task and calculates the information loss caused by its removal. A local LLM further analyzes boundary values and cross-sentence relations to determine the task value of each span. Based on task value and privacy risk, the framework applies retention, generalization, masking, or deletion to produce an initial sanitized prompt. TASK preserves more task-relevant information than existing LLM Context Aware methods. To mitigate residual disclosure, we further introduce an iterative security mechanism that evaluates the sanitized prompt using sensitive-value recovery and attribute-inference attacks. We compare our method with three existing privacy-preserving methods across document summarization, narrative question answering, and long-context question answering. The average task-quality and security scores improve by 23.72% and 8.64%, respectively.

Our main contributions are as follows: (1) We show that preserving output quality requires considering not only the sensitivity of information in a prompt but also its semantic contribution to the current task. (2) Based on this insight, we present TASK, a task-aware iterative prompt sanitization framework that combines task-relevance assessment with iterative security evaluation. (3) We evaluate TASK on long-context question-answering tasks and across multiple domains. The results show average improvements of 23.72% in task quality and 8.64% in security, demonstrating its effectiveness across different domains and task settings.

2 Related Work

As large language models such as GPT (OpenAI et al., 2024; Brown et al., 2020), Gemini (Team et al., 2025; Gemini Team et al., 2024), and Qwen3 (Team, 2025) continue to improve in language understanding and generation, users increasingly access these models through cloud-based APIs to complete complex tasks without local deployment (Zhu et al., 2024; Zhan et al., 2026). However, API-based inference requires users to transmit their prompts to the cloud, placing the data outside their local control. Cloud services may not only access sensitive information explicitly contained in prompts but also recover sensitive values or infer private attributes from contextual cues. Prompt privacy protection has therefore become an important concern in cloud-based LLM applications (Staab et al., 2024; Yoon et al., 2026). Existing studies addressing this problem can be broadly categorized into Prompt Redaction, Prompt Obfuscation, and LLM Context Aware.

2.1 Prompt Redaction

To prevent the direct disclosure of sensitive information, several methods (Kan et al., 2023; Serenari & Lee, 2025; Chen et al., 2023; Kleinberg et al., 2022; Liu et al., 2023) detect explicit identifiers and mask or replace them before prompts are transmitted. However, these methods primarily determine transformations according to predefined sensitive categories and cannot distinguish task-required facts from task-irrelevant identifiers. Even more comprehensive approaches, such as the framework proposed by Kan et al. (2023), which combines text anonymization, sensitive-information filtering, and response restoration, do not selectively preserve sensitive information according to its role in the current task. Consequently, Prompt Redaction may remove task-critical information in practical applications, leading to substantial degradation in LLM output quality.

2.2 Prompt Obfuscation

To reduce the direct exposure of original sensitive content, some methods (Lin et al., 2025b; Zhou et al., 2022; 2023; Pape et al., 2025; Lin et al., 2025a) obfuscate prompts by transforming tokens, embeddings, or symbolic representations. However, their privacy protection and task utility depend heavily on the perturbation configuration, making them difficult to adapt to the semantic and factual requirements of different tasks. Consequently, insufficient perturbation may leave sensitive information exposed, whereas excessive perturbation may destroy task-relevant entities, numerical values, and semantic relations, thereby substantially degrading model output quality.

2.3 LLM Context Aware

To improve the contextual awareness of privacy protection, several methods (Shen et al., 2026; Siyan et al., 2025; Yang et al., 2025; Pasch & Cha, 2025; Miranda et al., 2025) employ LLMs for privacy detection, prompt rewriting, or local–remote collaboration. However, these approaches generally determine which content to transform according to its context, without accurately assessing the semantic contribution of each sensitive span to the current task. As shown in Figure 1, although recent methods such as IntentAnony (Shen et al., 2026) attempt to improve task quality by selecting anonymization targets according to contextual intent, such intent does not necessarily reflect the specific requirements of the current task. Consequently, these methods may still fail to fully satisfy users' task requirements. In contrast, TASK addresses this limitation through two mechanisms: it evaluates the task value of sensitive spans based on semantic relevance and employs iterative security refinement to identify and mitigate residual risks, thereby improving privacy protection while preserving task-required information.

3 Method

3.1 Challenges

Challenge 1: Determining the Task Role of Sensitive Spans The quality of downstream responses depends strongly on information in the prompt that is relevant to the current task. In complex prompts, task-relevant information may be distributed across paragraphs, and sensitive spans

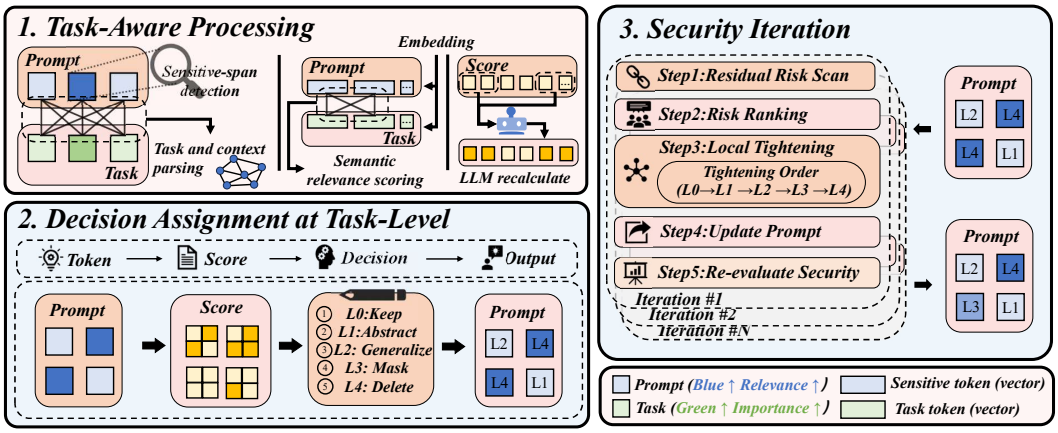


Figure 2: The design of TASK.(1) Task-aware processing identifies sensitive spans in the original prompt, parses the task, and evaluates the role of each span using embedding-based relevance scores and LLM review. (2) Task-level decision assignment maps each span to a sanitization level from $L0$ (keep) to $L4$ (delete), producing an initial sanitized prompt. (3) Security iteration detects residual disclosure risks through sensitive-value recovery and attribute-inference attacks and progressively strengthens selected sanitization decisions until the privacy requirement is satisfied.

may jointly support task semantics through entity, temporal, causal, and coreference relations. Accurately identifying and preserving this information is necessary to maintain the task objective, key facts, and output requirements, so that the sanitized prompt can still support a high-quality response.

Challenge 2: Mitigating Residual Risk while Preserving Task Semantics Task-relevant sanitization aims to preserve the semantic information needed to complete the task. It may therefore retain some highly task-relevant sensitive spans or apply weaker transformations, leaving sensitive cues in the final prompt and reducing privacy protection. The system must consequently evaluate each sanitization decision to determine whether the final prompt meets a predefined safety threshold. Under a limited number of evaluations, it must identify the decisions most likely to cause residual disclosure and determine whether they need further adjustment.

3.2 Framework Overview

As shown in Figure 2, given an original prompt P and a user-specified task T, TASK produces a sanitized prompt $\widehat{P}$. All sensitive-information analysis and sanitization decisions are performed locally. The framework consists of three stages.

Task-aware processing. We first identify candidate sensitive spans in P using domain lexicons and structured pattern matching, and record their types and positions. In parallel, we parse T to obtain the task objective, required facts, and output constraints. For each sensitive span, we extract its local context and organize related spans into privacy evidence chains. We then evaluate the task relevance of the span and its context, as well as the information loss caused by masking the span. An LLM further reviews borderline cases and cross-sentence relations to determine the task role of each sensitive span, thereby addressing Challenge 1.

Task-level decision assignment. Based on the task relevance, contextual risk, and task role of each sensitive span, the framework assigns an initial sanitization level. The available actions range from retaining the original value (L_0), abstraction (L_1), and generalization (L_2), to masking (L_3) and deletion (L_4). The sanitization strength increases from L_0 to L_4. Task-critical spans are assigned weaker transformations when possible to preserve task-required information, whereas task-irrelevant or high-risk spans receive stronger transformations. Applying these decisions produces the initial sanitized prompt.

Security iteration. The framework evaluates the initial sanitized prompt using sensitive-value recovery and attribute-inference attacks to identify residual disclosure risks. It ranks the sensitive spans associated with these risks and locally tightens their sanitization levels according to the order $L0 \rightarrow L1 \rightarrow L2 \rightarrow L3 \rightarrow L4$. After each update, the framework re-evaluates the prompt and repeats this process until the privacy requirement is satisfied. Only a prompt that passes the final safety gate is released to the downstream model, thereby addressing Challenge 2.

3.3 Task-Aware Processing

3.3.1 Candidate Detection and Parsing

Candidate detection takes the original prompt as input and produces a set of potentially sensitive spans. The detector integrates three sources of evidence. Structured patterns identify explicit values such as telephone numbers, account identifiers, and passwords. Domain lexicons and entity patterns cover names, organizations, addresses, and domain-specific information. Contextual scanning retains expressions with identifying capability.

Since sensitive spans of the same type may play different roles across tasks, we first parse the user task into a structured representation consisting of the task objective, key facts, and constraints on the output format and scope. Each candidate sensitive span is then jointly analyzed with this task representation to determine its relevance to the task objective, contribution to key facts, and impact on output constraints, providing a unified basis for subsequent task-relevance assessment and sanitization decisions.

3.3.2 Task-Semantic Scoring

For each sensitive span, we extract its local context and construct a masked counterpart by replacing the span with its sensitive-type marker. The task-semantic score combines three component scores: the direct relevance between the sensitive span and the task description, the relevance between the original context and the task, and the change in task relevance after masking the span.

Direct relevance indicates whether the sensitive span itself contributes to the current task, while context relevance measures whether its surrounding context supports the task. The change in relevance before and after masking estimates the resulting loss of task information. These three component scores are combined into the final task-semantic score. A higher score indicates that the span is more important for completing the current task. The detailed formulation is provided in Appendix A.

3.3.3 LLM-Based Contextual Adjudication

Within a fixed review budget, the system selects a subset of candidate spans according to their sensitive-information types and submits them to a local LLM for adjudication. For each candidate span, the LLM receives the task description, span text, detected sensitive type, task constraints, and local context. It determines whether the span is a valid sensitive-information candidate and returns its task role and recommended action. For event candidates, the LLM further determines whether the event contains sensitive facts required by the task and identifies the factual content that should be preserved. This review process is used to refine the initial BGE-based decisions. Valid sensitive candidates are then retained, generalized, or deleted according to the review results.

3.4 Task-level decision assignment

We formulate selective sanitization as task-conditioned exposure control. The system determines how much information a candidate may retain based on its task role, context, and privacy-inference risk. We define the processing levels as

$$\mathcal{L} = \{L_0, L_1, L_2, L_3, L_4\}. \tag{1}$$

Here, L_0 retains the original value; L_1 abstracts the sensitive value while preserving task-relevant semantics; L_2 generalizes the value into a broader category or range; L_3 replaces it with a mask; and L_4 deletes it. Sanitization strength increases progressively from L_0 to L_4. If no feasible transformation satisfies the release criteria within the maximum number of refinement iterations, the prompt is withheld from automatic release and sent for human review.

The system determines the processing level for each sensitive attribute from the candidate's position, associated entities and events, contextual evidence, and current task. The decision depends on its task role and privacy risk. When a candidate contains multiple sensitive attributes, the strictest applicable requirement governs its transformation. Furthermore, since combinations of different sensitive spans may still support joint privacy inference, the system further limits the identifying cues retained in the sanitized prompt for privacy evidence chains spanning multiple candidates.

The selected transformations are applied to the original prompt to produce an initial sanitized prompt. Before security refinement, the system checks whether this prompt preserves the task objective, required facts, and sufficient information to complete the task.

3.5 Security iteration

Task-aware sanitization prioritizes task semantics and may therefore leave privacy cues in the initial prompt. At each iteration, the system evaluates three sources of residual risk: direct exposure of sensitive values, recovery of sensitive values from context, and inference of sensitive attributes from combinations of spans. Their weighted aggregate is used to compute the privacy score $S(P^{(t)})$ of the current prompt. The refinement terminates when $S(P^{(t)}) \geq \tau_S$ and the task structure and required information are preserved, where τ_S denotes the safety release threshold; otherwise, the system performs local refinement. We set $\tau_S = 0.90$ in our experiments, with the threshold-selection analysis provided in Appendix B.2.

For candidates associated with residual risk, the system considers feasible transformations stricter than their current ones. It estimates the privacy-risk reduction and task-information loss of each candidate transformation, with task loss determined by the candidate's task value and the impact of the transformation. The next candidate and transformation are selected by

$$(i^\star, m^\star) = \arg \max_{\substack{u_i \in \mathcal{F}^{(t)} \\ m \in \mathcal{M}^{(t)}(u_i)}} \frac{\Delta R_{i,m}^{(t)}}{\Delta Q_{i,m}^{(t)} + \varepsilon}, \tag{2}$$

where $\mathcal{F}^{(t)}$ is the set of candidates associated with residual risk, and $\mathcal{M}^{(t)}(u_i)$ is the set of feasible stricter transformations for u_i. The terms $\Delta R_{i,m}^{(t)}$ and $\Delta Q_{i,m}^{(t)}$ denote risk reduction and task-information loss, respectively, while $\varepsilon = 10^{-6}$ prevents division by zero. The criterion selects the transformation with the greatest risk reduction per unit of task loss and re-evaluates the prompt after each update. Detailed definitions are provided in Appendix B.1.

For task-critical content, the system progressively strengthens the sanitization strategy through feasible transformations from L_0 to L_4, while avoiding unnecessary loss of task-relevant information. After each update, the system recomputes the privacy score and re-evaluates task preservation until the release criterion is satisfied or the stopping condition is reached.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method in three settings: narrative question answering, long-context question answering, and domain-specific document summarization. These settings test whether a sanitization method can preserve facts needed to answer questions and retain evidence distributed over long contexts. For narrative question answering, we use NarrativeQA (Kočiský et al., 2018), which contains narrative texts and corresponding questions. For long-context question answering, we use InfiniteBench (Zhang et al., 2024), a benchmark designed to evaluate information retrieval and reasoning over long inputs. For document summarization, we use CNN/DailyMail (See et al., 2017). We retrieve and adapt articles from four domains—education, finance, law, and healthcare—to construct prompts containing both task-relevant and task-irrelevant sensitive information. Table 3 summarizes the resulting evaluation data.

Baselines. We compare our method with Text Redaction (Aahill, 2023), IntentAnony (Shen et al., 2026), and Prompt Obfuscation (Pape et al., 2025). Text Redaction detects sensitive spans using

manually specified rules or a named-entity recognizer and then removes each detected span or replaces it with a fixed mask. Its transformations depend on predefined sensitive-entity categories and do not explicitly estimate whether a span is required by the downstream task. IntentAnony identifies the communicative intent of the input and constructs privacy-inference evidence chains from contextual cues. It assigns an exposure budget to each privacy attribute conditioned on the inferred intent and suppresses evidence unrelated to that intent while preserving content needed to express it. A constrained LLM rewrite produces the anonymized prompt at inference time. Prompt Obfuscation provides hard- and soft-prompt variants. The hard-prompt variant optimizes a functionally equivalent but less readable discrete token sequence, whereas the soft-prompt variant directly optimizes continuous prompt embeddings. The obfuscated representation is then supplied to the downstream model together with the user query.

Evaluation Metrics. We use a privacy score and a task-quality score as our primary metrics. Both range from 0 to 1, with higher values indicating better performance. The detailed formulation is provided in Appendix C. We evaluate residual privacy risk using sensitive-value recovery and sensitive-attribute inference attacks. The privacy score is defined as

$$S = 1 - (0.5R_{\text{rec}} + 0.5R_{\text{attr}}), \tag{3}$$

where R_{rec} and R_{attr} denote sensitive-value recovery risk and attribute-inference risk, respectively. Both risks are calculated from the agreement between attack predictions and the original sensitive information, together with attack confidence and sensitivity-based weighting. A higher S indicates greater resistance to privacy recovery and inference attacks. We further evaluate task quality by comparing the downstream model's response A to the original prompt with its response A' to the sanitized prompt:

$$Q = 0.5F_{\text{tok}}(A, A') + 0.5Q_{\text{sem}}(T, A'), \tag{4}$$

where F_{tok} is the token-level F1 score between A and A', and Q_{sem} measures how completely A' satisfies the task requirements. The semantic evaluation considers content coverage, factual consistency, and constraint satisfaction, following the rubric-based paradigm of LLM-assisted evaluation (Shen et al., 2025). This metric captures both surface-level answer consistency and task completion, thereby measuring the quality loss introduced by prompt sanitization.

Models. We use bge-base-en-v1.5 Xiao et al. (2023) to encode sensitive spans, tasks, and local contexts, and a locally deployed `glm-4-9B` GLM et al. (2024) to perform boundary-case adjudication, privacy attacks, and task-semantic evaluation. In our subsequent experiments analyzing the relationship between model sensitivity and TASK performance, we further compare GLM-4-9B with Qwen3-32B (Team, 2025) and Llama-3-8B (Dubey et al., 2024). All methods are executed in the same hardware environment.

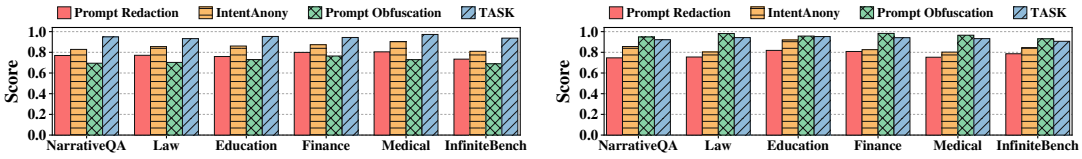
4.2 QUALITY–PRIVACY EVALUATION

Task Quality As shown in Figure 3a, Averaging quality scores across the four CNN/DailyMail domains, TASK improves over Prompt Redaction by 21.30%, over Prompt Obfuscation by 29.95%, and over IntentAnony by 8.82%. Uniform deletion, masking, or representation-level obfuscation can remove or distort entities, numerical values, and factual relations required by the task, whereas TASK explicitly estimates task relevance, local context, and masking impact for each sensitive span. On NarrativeQA and InfiniteBench, TASK improves quality over Prompt Redaction by 23.39% and 27.70%, over Prompt Obfuscation by 36.68% and 36.02%, and over IntentAnony by 14.63% and 15.74%, respectively. These long-context tasks depend on entities, events, and relations distributed throughout the input; conditioning transformation strength on the role of each span better preserves the evidence needed to answer the question. Overall, as shown in Table 1, TASK improves task quality over Prompt Redaction, IntentAnony, and Prompt Obfuscation by 24.06%, 12.97%, and 34.14%, respectively, showing that task-aware sanitization reduces the utility loss caused by uniform privacy transformations. The detailed information is provided in Appendix D.1.

Privacy Protection As shown in Figure 3b, our method improves the average security score over IntentAnony by 9.12%. The per-setting improvements range from 3.60% to 17.29%, with gains of 17.29%, 14.03%, and 16.23% on CNN-Law, CNN-Finance, and CNN-Medical, respectively, and gains of 7.70% and 7.23% on NarrativeQA and InfiniteBench. IntentAnony selects anonymization

Table 1: Average quality score and privacy score across three evaluation settings. Relative changes compare TASK with each baseline.

Metric	Prompt Redaction	IntentAnony	Prompt Obfuscation	TASK
Average quality score	0.7623	0.8371	0.7051	0.9457
Average privacy score	0.7721	0.8463	0.9503	0.9235
TASK vs. baseline: quality (%)	+24.06	+12.97	+34.14	–
TASK vs. baseline: privacy (%)	+19.61	+9.12	-2.82	–



(a) Quality Score.

(b) Privacy Score.

Figure 3: Performance comparison of TASK against three baseline methods across six evaluation groups. Subplot (a) presents task-quality scores, and subplot (b) presents privacy scores.

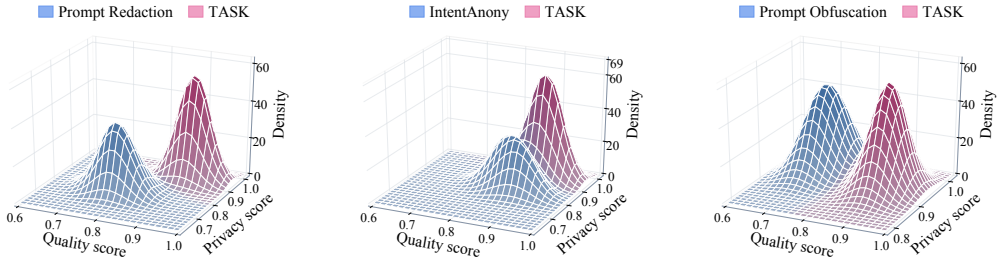


Figure 4: Distribution of task-quality and privacy scores for TASK and baseline methods. TASK produces a distribution more concentrated in higher-value regions than the baselines, indicating its stronger ability to preserve task semantics while protecting privacy.

targets using communicative intent and contextual evidence, but does not use the same attack-guided refinement procedure. By applying recovery and attribute-inference attacks after initial sanitization, our method can identify and tighten decisions that continue to expose sensitive cues.

Relative to Text Redaction, our method improves average security by 19.61% and quality by 24.06%. Relative to Prompt Obfuscation, our method improves quality by 34.14%, while its security score is 2.82% lower. These comparisons demonstrate that TASK preserves task-relevant information while satisfying the predefined privacy threshold under the evaluated attack setting, thereby improving the quality of downstream task outputs.

4.3 Quality–Privacy Distribution Analysis.

The joint distribution of task-quality and privacy scores reflects the overall performance of a sanitization method in preserving task semantics while reducing privacy risks. Examining score distributions across six evaluation groups helps assess the stability of each method across tasks and domains. Figure 4 compares the quality–privacy distributions of TASK and three baseline methods. Compared with Prompt Redaction and IntentAnony, TASK exhibits density peaks more concentrated in regions with high task quality and privacy, indicating that it more consistently preserves task-required information while reducing residual risk. Compared with Prompt Obfuscation, TASK shifts substantially toward higher task-quality regions while maintaining high privacy scores, demonstrating a better privacy–utility balance.

Table 2: Comparison of quality score (Q) and privacy score (S) achieved by TASK with GLM-4-9B, Qwen3-32B, and Llama-3-8B.

Dataset	Domain	GLM-4-9B		Qwen3-32B		Llama-3-8B	
		Q	S	Q	S	Q	S
CNN/DailyMail	Education	0.9034	0.9523	0.9267	0.9372	0.9068	0.9212
	Law	0.8824	0.9423	0.9195	0.9483	0.8895	0.9411
	Finance	0.9135	0.9414	0.9370	0.9638	0.8906	0.9403
	Medical	0.9727	0.9323	0.9523	0.9512	0.9305	0.9391
NarrativeQA	–	0.8894	0.9218	0.9074	0.9229	0.8844	0.9309
InfiniteBench	–	0.8673	0.9066	0.8855	0.9147	0.8744	0.9234
Average Score.		0.8916	0.9235	0.9089	0.9292	0.8877	0.9299
Other models vs. GLM (%)		–	–	+1.95	+0.62	-0.43	+0.69

Dataset	**Samples**	**Avg. tokens**
NarrativeQA	300	2,071
InfiniteBench	300	1,871
CNN/DailyMail	200	2,568

Table 3: Evaluation datasets (NarrativeQA, InfiniteBench, and CNN/DailyMail), including the number of samples and average token count.

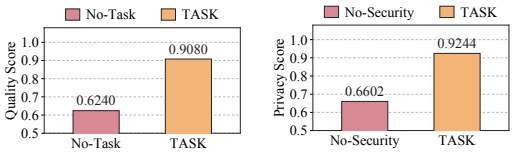


Figure 5: Component ablation results of TASK. Left: TASK improves quality by 45.56% over non-task-aware processing. Right: TASK improves security by 40.02% over non-security processing.

4.4 Cross-Model Sensitivity

We evaluate TASK across six evaluation groups: CNN/DailyMail in the education, law, finance, and medical domains, NarrativeQA, and InfiniteBench. We use GLM-4-9B (GLM et al., 2024), Qwen3-32B (Team, 2025), and Llama-3-8B (Dubey et al., 2024), while keeping all other procedures fixed. As shown in Table 2, the average task-quality and privacy scores vary by approximately 0.4%–2% across the three models. These results indicate that TASK remains robust to model choice and can consistently maintain high task quality and privacy protection when using different models.

4.5 Ablation Study

We remove task-aware processing and security refinement separately to evaluate their respective contributions. Figure 5 reports results from the ablation configuration. Removing task-aware processing reduces the quality score from 0.9080 to 0.6240, indicating that transformations based only on sensitive types discard task-relevant facts and relations. Removing security refinement reduces the security score from 0.9244 to 0.6602, showing that initial selective sanitization can leave residual or compositional risks. Together, these ablations suggest that task-aware processing primarily preserves utility, whereas attack-guided refinement primarily mitigates residual disclosure. The detailed formulation is provided in Appendix D.2.

5 Conclusion

This paper presents TASK, a task-aware iterative prompt sanitization framework for cloud-based LLM inference. Its design is motivated by two observations: sensitive spans of the same type can play different semantic roles in the current task, and residual privacy risks after initial sanitization are concentrated in a small number of retained or weakly transformed spans and their associated evidence. Based on these observations, TASK combines task-relevance assessment for sensitive spans with iterative privacy-risk assessment. Evaluation results show that, compared with three existing methods, TASK improves task quality and privacy protection by 23.72% and 8.64% on average, respectively.

AI USE STATEMENT

In this work, we used Large Language Models (LLMs) as auxiliary tools to improve research efficiency and manuscript quality. All content generated or modified with the assistance of LLMs was reviewed and verified by the authors to ensure compliance with the principles of academic integrity. LLMs were used solely as research-support tools, and the authors take full responsibility for the research methodology, experimental results, claims, citations, and final conclusions presented in this work. A detailed description of the specific uses of these models is provided below:

LANGUAGE POLISHING.

We used LLMs to assist with language polishing and text editing, including improvements to grammar, clarity, academic writing style, and the organization of selected passages. These modifications were intended primarily to improve the readability and presentation quality of the manuscript. All AI-assisted revisions were manually reviewed by the authors to ensure that they did not alter the original research ideas, technical methodology, or experimental conclusions.

LITERATURE ORGANIZATION.

During the research and manuscript preparation process, we used LLMs to assist in summarizing established research concepts and organizing related work, which helped identify potential research gaps. All related work, literature citations, and corresponding discussions included in the final manuscript were independently checked and verified by the authors.

REFERENCES

Aahill. What is azure ai language - azure ai services, 2023. URL https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview. Accessed on Jan 12, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

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*, 2023. URL https://arxiv.org/abs/2309.03057.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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, 2025. URL https://arxiv.org/abs/2407.02956.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In Elena Volodina, David Alfter, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, and Xuan-Son Vu (eds.), *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pp. 64–75, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.caldpseudo-1.8. URL https://aclanthology.org/2024.caldpseudo-1.8/.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, and et al. Wu, Yu. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Shilong Hou, Ruilin Shang, Zi Long, Xianghua Fu, and Yin Chen. A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation. *arXiv preprint arXiv:2502.15233*, 2025. URL https://arxiv.org/abs/2502.15233.

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, 2023. URL https://arxiv.org/abs/2306.08223.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. Textwash – automated open-source text anonymisation, 2022. URL https://arxiv.org/abs/2208.13081.

Tomáš Kočiský, 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. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Sam Lin, Wenyue Hua, Zhenting Wang, Mingyu Jin, Lizhou Fan, and Yongfeng Zhang. EmojiPrompt: Generative prompt obfuscation for privacy-preserving communication with cloud-based LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12342–12361, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.614. URL https://aclanthology.org/2025.naacl-long.614/.

Yu Lin, Ruining Yang, Yunlong Mao, Qizhi Zhang, Jue Hong, Quanwei Cai, Ye Wu, Huiqi Liu, Zhiyu Chen, Bing Duan, and Sheng Zhong. ObfusLM: Privacy-preserving language model service against embedding inversion attacks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 1160–1174. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.58. URL https://aclanthology.org/2025.acl-long.58/.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. DeID-GPT: Zero-shot medical text de-identification by GPT-4, 2023. URL https://arxiv.org/abs/2303.11032.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. Preserving privacy in large language models: A survey on current threats and solutions. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ss9MTTN7OL.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and et al. Shyamal Anadkat. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

David Pape, Sina Mavali, Thorsten Eisenhofer, and Lea Schönherr. Prompt obfuscation for large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2323–2342, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/pape.

Stefan Pasch and Min Chul Cha. Balancing privacy and utility in personal LLM writing tasks: An automated pipeline for evaluating anonymizations. In Ivan Habernal, Sepideh Ghanavati, Vijayanta Jain, Timour Igamberdiev, and Shomir Wilson (eds.), *Proceedings of the Sixth Workshop on Privacy in Natural Language Processing*, pp. 32–41, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-246-6. doi: 10.18653/v1/2025.privatenlp-main.3. URL https://aclanthology.org/2025.privatenlp-main.3/.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Jayden Serenari and Stephen Lee. Semantically-aware llm agent to enhance privacy in conversational ai services. *arXiv preprint arXiv:2510.27016*, 2025. URL https://arxiv.org/abs/2510.27016.

Weihao Shen, Yaxin Xu, Shuang Li, Wei Chen, Yuqin Lan, Meng Yuan, and Fuzhen Zhuang. You only anonymize what is not intent-relevant: Suppressing non-intent privacy evidence, 2026. URL https://arxiv.org/abs/2601.04265.

Xicong Shen, Yang Liu, Yi Liu, Peiran Wang, Huiqi Liu, Jue Hong, Bing Duan, Zirui Huang, Yunlong Mao, Ye Wu, and Sheng Zhong. Sap: Privacy-preserving fine-tuning on language models with split-and-privatize framework. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 502–510. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/57. URL https://doi.org/10.24963/ijcai.2025/57. Main Track.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy preservation from Internet-based and local language model ensembles. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3390, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.173. URL https://aclanthology.org/2025.naacl-long.173/.

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*, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, and et al. Amelia Glaese. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28922–28941, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1404. URL https://aclanthology.org/2025.acl-long.1404/.

Jeongho Yoon, Chanhee Park, Yongchan Chun, Hyeonseok Moon, and Heuiseok Lim. Towards privacy-preserving large language model: Text-free inference through alignment and adaptation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25961–25981. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.1191/.

Junfei Zhan, Haoxun Shen, Zheng Lin, and Tengjiao He. PRISM: Privacy-aware routing for adaptive cloud-edge LLM inference via semantic sketch collaboration. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, pp. 28150–28158. AAAI Press, 2026. URL https://arxiv.org/abs/2511.22788.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.814.

Xin Zhou, Jinzhu Lu, Tao Gui, Ruotian Ma, Zichu Fei, Yuran Wang, Yong Ding, Yibo Cheung, Qi Zhang, and Xuanjing Huang. TextFusion: Privacy-preserving pre-trained model inference via token fusion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8360–8371. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.572. URL https://aclanthology.org/2022.emnlp-main.572/.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5459–5473. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.337. URL https://aclanthology.org/2023.findings-acl.337/.

Youxiang Zhu, Ning Gao, Xiaohui Liang, and Honggang Zhang. Exploiting privacy preserving prompt techniques for online large language model usage. In *2024 IEEE Global Communications Conference*, pp. 4304–4309. IEEE, 2024. doi: 10.1109/GLOBECOM52923.2024.10901426. URL https://doi.org/10.1109/GLOBECOM52923.2024.10901426.

Tymon Lesław Żarski and Artur Janicki. Enhancing privacy while preserving context in text transformations by large language models. *Information*, 16(1), 2025. ISSN 2078-2489. doi: 10.3390/info16010049. URL https://www.mdpi.com/2078-2489/16/1/49.

Algorithm 1 Task-Semantic Scoring and Local Adjudication

Require: Original prompt P, task T, candidate set $\mathcal{U}$
Ensure: Task-semantic scores $\{q_i\}$ and reviewed candidate decisions $\mathcal{D}$
Parse T into task objective, target object, required facts, and output constraints
$h_T \leftarrow f_{\mathrm{BGE}}(T)$
for all $u_i = (x_i, \tau_i, p_i) \in \mathcal{U}$ **do**
    Extract local context c_i around x_i
    Construct masked context c_i^- by replacing x_i with a type marker
    Encode $\mathrm{Desc}(x_i, \tau_i, c_i)$, c_i, and c_i^-
    Compute s_i^{dir}, s_i^{ctx}, s_i^-, and d_i^{del}
    Compute q_i using Eq. equation 7
end for
Select boundary candidates according to score uncertainty, d_i^{del}, and cross-sentence evidence
for all selected candidate u_i **do**
    Query the local LLM with T, x_i, τ_i, task constraints, and c_i
    Obtain candidate validity, task role, and recommended action
    Update the candidate decision if the local review contradicts the initial assessment
end for
return $\{q_i\}$ and $\mathcal{D}$

A DETAILED FORMULATION OF TASK-SEMANTIC SCORING

For each candidate sensitive span $u_i = (x_i, \tau_i, p_i)$, where x_i, τ_i, and p_i denote its text, detected type, and position, respectively, we extract a local context c_i and construct a masked counterpart c_i^-. We encode the task description, candidate description, original context, and masked context using the semantic encoder:

$$h_T = f_{\mathrm{BGE}}(T), \quad h_i^{\mathrm{dir}} = f_{\mathrm{BGE}}(\mathrm{Desc}(x_i, \tau_i, c_i)), \quad h_i^{\mathrm{ctx}} = f_{\mathrm{BGE}}(c_i), \quad h_i^- = f_{\mathrm{BGE}}(c_i^-). \tag{5}$$

We compute the direct task relevance and contextual task relevance as

$$s_i^{\mathrm{dir}} = \frac{1 + \cos(h_i^{\mathrm{dir}}, h_T)}{2}, \qquad s_i^{\mathrm{ctx}} = \frac{1 + \cos(h_i^{\mathrm{ctx}}, h_T)}{2}, \qquad s_i^- = \frac{1 + \cos(h_i^-, h_T)}{2}. \tag{6}$$

The difference $s_i^{\mathrm{ctx}} - s_i^-$ measures the task-relevant information lost after masking x_i. The final task-semantic score is

$$q_i = 0.35 s_i^{\mathrm{dir}} + 0.30 s_i^{\mathrm{ctx}} + 0.35 \max(0, s_i^{\mathrm{ctx}} - s_i^-). \tag{7}$$

We also compute the local deletion impact

$$d_i^{\mathrm{del}} = 1 - \cos(h_i^{\mathrm{ctx}}, h_i^-), \tag{8}$$

which measures the semantic change caused by removing the candidate from its local context. This value is used as auxiliary evidence during local LLM adjudication, particularly for boundary cases and cross-sentence dependencies.

B DETAILED FORMULATION OF ATTACK-GUIDED SECURITY REFINEMENT

B.1 FORMULA ANALYSIS

Given the sanitized prompt $P^{(t)}$ at refinement round t, the system evaluates residual disclosure through sensitive-value recovery and attribute-inference attacks. For each original sensitive span x_i, the recovery attacker generates a candidate set $\mathcal{C}_{\mathrm{rec}}^{(t)}$, and the attribute attacker generates a candidate set $\mathcal{C}_{\mathrm{attr}}^{(t)}$. The span-level risks are

$$r_i^{(t)} = \max_{c \in \mathcal{C}_{\mathrm{rec}}^{(t)}} \mathrm{Match}_{\mathrm{rec}}(x_i, c)\, \mathrm{Conf}(c), \qquad a_i^{(t)} = \max_{c \in \mathcal{C}_{\mathrm{attr}}^{(t)}} \mathrm{Match}_{\mathrm{attr}}(x_i, c)\, \mathrm{Conf}(c). \tag{9}$$

Each candidate is assigned a detector-derived sensitivity score $s_i \in [0, 1]$ during sensitive-span detection. High-confidence structured identifiers receive a fixed high score, semantic-event candidates use their BGE prototype similarity, and other candidates retain the score produced by their detector. This score is fixed before refinement and is independent of downstream answers and attack outcomes. We set $\omega_i = \max(0.1, s_i)$ and aggregate the attack risks as

$$R_{\text{rec}}^{(t)} = \frac{\sum_{i=1}^{n} \omega_i r_i^{(t)}}{\sum_{i=1}^{n} \omega_i}, \qquad R_{\text{attr}}^{(t)} = \frac{\sum_{i \in \mathcal{A}} \omega_i a_i^{(t)}}{\sum_{i \in \mathcal{A}} \omega_i}, \tag{10}$$

where $\mathcal{A}$ contains candidates eligible for attribute-inference evaluation. We set $R_{\text{attr}}^{(t)} = 0$ when $\mathcal{A}$ is empty. The residual risk and privacy score are

$$R(P^{(t)}) = 0.5R_{\text{rec}}^{(t)} + 0.5R_{\text{attr}}^{(t)}, \qquad S(P^{(t)}) = 1 - R(P^{(t)}). \tag{11}$$

When $S(P^{(t)}) < \tau_S$, the system considers only stricter feasible transformations. The sanitization levels are ordered as

$$L_0 \text{ (keep original value)} \rightarrow L_1 \text{ (abstract)} \rightarrow L_2 \text{ (generalize)} \rightarrow L_3 \text{ (mask)} \rightarrow L_4 \text{ (delete)}. \tag{12}$$

Let ℓ_i denote the current level of candidate u_i. The feasible action set $\mathcal{M}^{(t)}(u_i)$ contains only transformations m whose level is stricter than ℓ_i and that do not violate explicit task constraints.

For every feasible candidate-action pair (u_i, m), we temporarily apply m to obtain a counterfactual prompt $P_{i,m}^{(t)}$. The same attack protocol is then applied to compute its residual risk. The privacy gain is defined as

$$\Delta R_{i,m}^{(t)} = R(P^{(t)}) - R(P_{i,m}^{(t)}). \tag{13}$$

To estimate task-information loss without executing the downstream task for every candidate action, we reuse the task-semantic score q_i. Let $c_i^{(t)}$ denote the current local context of u_i in $P^{(t)}$, and let $c_{i,m}^{(t)}$ denote the corresponding context after applying transformation m.

$$d_{i,m}^{(t)} = 1 - \cos\left(f_{\text{BGE}}(c_i^{(t)}), f_{\text{BGE}}(c_{i,m}^{(t)})\right). \tag{14}$$

The estimated task-information loss is

$$\Delta Q_{i,m}^{(t)} = q_i d_{i,m}^{(t)}. \tag{15}$$

Thus, transformations applied to task-critical spans or causing substantial contextual changes receive larger penalties.

The next refinement action is selected by

$$(i^\star, m^\star) = \arg \max_{\substack{u_i \in \mathcal{F}^{(t)} \\ m \in \mathcal{M}^{(t)}(u_i)}} \frac{\Delta R_{i,m}^{(t)}}{\Delta Q_{i,m}^{(t)} + \varepsilon}, \tag{16}$$

where $\mathcal{F}^{(t)}$ contains candidates implicated by nonzero recovery or attribute-inference risk. We set $\varepsilon = 10^{-6}$ solely for numerical stability. The resulting ratio prioritizes transformations that achieve larger reductions in residual privacy risk while introducing smaller losses of task-relevant information.

B.2 Privacy Threshold

We study the effect of the safety release threshold τ_S on the privacy–utility trade-off using a held-out development set. Specifically, we evaluate $\tau_S \in \{0.80, 0.85, 0.90\}$ and run the complete sanitization and security-refinement procedure under each setting. For each threshold, we report the resulting task-quality score Q and privacy score S across the four domains. The development samples used for threshold selection are disjoint from the test samples used in the main evaluation.

As shown in Table 4, increasing τ_S generally strengthens privacy protection while maintaining high task quality. At $\tau_S = 0.80$, the privacy scores range from 0.8395 to 0.8784, whereas increasing

Algorithm 2 Attack-Guided Security Refinement

Require: Initial prompt $P^{(0)}$, candidate decisions $\mathcal{D}$, task scores $\{q_i\}$, threshold τ_S, maximum iterations $T_{\max}$
Ensure: Released prompt $\widehat{P}$ or a human-review request

for $t = 0$ to $T_{\max}$ **do**
  Run recovery and attribute-inference attacks on $P^{(t)}$
  Compute $R(P^{(t)})$ and $S(P^{(t)})$
  if $S(P^{(t)}) \geq \tau_S$ and task constraints are satisfied **then**
    return $P^{(t)}$
  end if
  Construct $\mathcal{F}^{(t)}$ from candidates implicated by attack outputs
  for all $u_i \in \mathcal{F}^{(t)}$ **do**
    for all $m \in \mathcal{M}^{(t)}(u_i)$ **do**
      Construct counterfactual prompt $P_{i,m}^{(t)}$
      Run the same attacks on $P_{i,m}^{(t)}$ and compute $R(P_{i,m}^{(t)})$
      Compute $\Delta R_{i,m}^{(t)}$ using Eq. equation 13
      Compute $\Delta Q_{i,m}^{(t)}$ using Eq. equation 15
    end for
  end for
  if no feasible action yields $\Delta R_{i,m}^{(t)} > 0$ **then**
    return HUMANREVIEW
  end if
  Select $(i^\star, m^\star)$ using Eq. equation 16
  Apply $m^\star$ to obtain $P^{(t+1)}$
end for
return HUMANREVIEW

Table 4: Effect of the privacy threshold τ_S on task quality (Q) and privacy score (S) across four domains.

Privacy Threshold	Education		Law		Finance		Medical	
	Q	S	Q	S	Q	S	Q	S
0.80	0.9626	0.8699	0.9337	0.8761	0.9775	0.8395	0.9819	0.8784
0.85	0.9592	0.8838	0.9419	0.8769	0.9583	0.8861	0.9629	0.8957
0.90	0.9534	0.9523	0.9324	0.9423	0.9435	0.9414	0.9727	0.9323

the threshold to 0.85 raises the privacy scores to 0.8769–0.8957. With $\tau_S = 0.90$, all four domains achieve privacy scores above 0.93, ranging from 0.9323 to 0.9523, while the task-quality scores remain between 0.9324 and 0.9727. These results indicate that a stricter release threshold leads the refinement process to further mitigate residual privacy cues, with only a limited reduction in task utility.

Based on the privacy–utility trade-off observed on the development set, we set $\tau_S = 0.90$ as the safety release threshold and keep it fixed for all subsequent test-set experiments. No test-set results are used for threshold selection.

C DETAILED FORMULATION OF EVALUATION METRICS

We use a privacy score and a task-quality score as our primary metrics. Both range from 0 to 1, with higher values indicating better performance.

Privacy score. Given a sanitized prompt P', the attacker attempts both sensitive-value recovery and sensitive-attribute inference. We define the privacy score as

$$S = 1 - (0.5R_{\text{rec}} + 0.5R_{\text{attr}}), \tag{17}$$

where R_{rec} and R_{attr} denote recovery risk and attribute-inference risk, respectively.

For each sensitive span in the original prompt, we compare the attack candidates with the original value. The maximum product of candidate-match score and attack confidence gives its span-level risk. We aggregate the two risks using sensitivity-based weights:

$$R_{\text{rec}} = \frac{\sum_{i=1}^{n} \omega_i r_i}{\sum_{i=1}^{n} \omega_i}, \qquad R_{\text{attr}} = \frac{\sum_{i \in \mathcal{A}} \omega_i a_i}{\sum_{i \in \mathcal{A}} \omega_i}. \tag{18}$$

Here, r_i and a_i are the span-level recovery and attribute-inference risks, and $\mathcal{A}$ indexes spans eligible for attribute-inference evaluation. The weight $\omega_i = \max(0.1, s_i)$ is determined by the span's sensitivity s_i; the lower bound prevents a zero weight. We set $R_{\text{attr}} = 0$ when $\mathcal{A}$ is empty.

Recovery matching considers normalized exact matches, substring matches, numerical overlap, and longest-common-subsequence similarity. Attribute matching considers value agreement, type agreement, and type-specific semantic rules. The attacker receives only P' and the target sensitive types. To avoid attacker–evaluator circularity, the attacker used for iterative security refinement is separated from the held-out attackers used for final privacy evaluation, whose outputs are never exposed to the refinement process.

Task-quality score. Let A be the answer generated from the original prompt and A' the answer produced by the same downstream model from the sanitized prompt. We measure utility as

$$Q = 0.5F_{\text{tok}}(A, A') + 0.5Q_{\text{sem}}(T, A'), \tag{19}$$

where $F_{\text{tok}} \in [0, 1]$ is the token-level F1 score between A and A', and $Q_{\text{sem}} \in [0, 1]$ measures how completely A' satisfies the task T. Token-level F1 is

$$F_{\text{tok}} = \frac{2PR}{P + R}, \tag{20}$$

where P and R are the token precision and recall of A' with respect to A.

The semantic evaluator decomposes T into $m \in [3, 8]$ independently assessable requirements. For each requirement j, it assigns a continuous completion score $q_j \in [0, 1]$ based on content coverage, factual consistency, and constraint satisfaction. A score near 1 indicates complete and correct satisfaction, an intermediate value indicates partial satisfaction, and a score near 0 indicates omission or contradiction. The aggregate semantic score is

$$Q_{\text{sem}} = \frac{\sum_{j=1}^{m} \lambda_j q_j}{\sum_{j=1}^{m} \lambda_j}, \tag{21}$$

where λ_j is the weight of requirement j. Factual contradictions reduce the corresponding q_j. All methods share the same task decomposition, requirement weights, evaluator prompt, and generation parameters. Combining F_{tok} with Q_{sem} reduces reliance on either lexical overlap or a single LLM judgment.

D DETAILED EXPERIMENTAL RESULTS AND ANALYSIS

D.1 DETAILED PERFORMANCE COMPARISON

Table 5: Task quality (Q) and privacy (S) across six evaluation groups. Macro averages give equal weight to CNN/DailyMail, NarrativeQA, and InfiniteBench.

Dataset	**Domain**	**Metric**	**Text Redaction**	**IntentAnony**	**Prompt Obfuscation**	**Ours**
CNN/DailyMail	Education	Q	0.7594	0.8608	0.7299	0.9534
		S	0.8186	0.9192	0.9571	0.9523
	Law	Q	0.7712	0.8553	0.7024	0.9324
		S	0.7547	0.8034	0.9814	0.9423
	Finance	Q	0.7993	0.8736	0.7643	0.9435
		S	0.8081	0.8256	0.9829	0.9414
	Medical	Q	0.8046	0.9040	0.7292	0.9727
		S	0.7525	0.8021	0.9653	0.9323
NarrativeQA	–	Q	0.7694	0.8282	0.6946	0.9494
		S	0.7465	0.8559	0.9483	0.9218
InfiniteBench	–	Q	0.734	0.8098	0.6891	0.9373
		S	0.7862	0.8455	0.9309	0.9066
Three-setting macro avg. (Q)			0.7623	0.8371	0.7051	0.9457
Three-setting macro avg. (S)			0.7721	0.8463	0.9503	0.9235
Ours vs. baseline: quality (%)			+24.06	+12.97	+34.14	–
Ours vs. baseline: privacy (%)			+19.61	+9.12	-2.82	–

D.2 ABLATION STUDY

Table 6: Ablation results and improvements of the full model over each variant.

Method	Quality	Security	Quality Improvement (%)	Security Improvement (%)
Non-task-aware Processing	0.6240	–	45.5611	–
Ours	0.9080	0.9244	–	–
Non-security Processing	–	0.6602	–	40.0182