Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33648

A Probabilistic Risk Scoring and Dynamic Mitigation Framework for Privacy Protection in LLM QA Applications

Sirui Zhang

Department of Computer Science and Technology, School of Computer Science and Technology, Dalian University of Technology, Dalian, China
2173878265@qq.com

Abstract: LLM based question answering systems have been used in healthcare, education, and customer support, and in these settings user prompts often contain names, health conditions, contact details, or other sensitive clues, which makes privacy protection at inference time hard to avoid, especially when the service is deployed as a black box. Many existing defenses still rely on retraining or relatively rigid filtering rules, so they do not adapt well when contextual sensitivity changes from one interaction to another. This paper proposes a privacy protection framework for LLM QA at the question answer pair level. The framework estimates exposure risk with a probabilistic graphical model, then adjusts mitigation strength through threshold based control. In the protection stage, sanitization, abstraction, and calibrated noise are used together, so stronger intervention can be applied to more sensitive inputs while ordinary interactions are affected less. Experiments on a SafetyBench derived dataset show that the framework can reduce privacy risk and still keep answer utility at a useful level, with relatively low computational overhead.

Keywords: Large Language Models, Privacy Protection, Question Answering, Risk Assessment, Privacy Utility Balance

1. Introduction

Large language models are now used more often in QA services for medical consultation, education, and customer support, and in these scenarios privacy risk tends to appear in ordinary interaction rather than only in training data, because users may directly type identity attributes, personal events, or health related details into their questions.

At the same time, many existing safeguards are not easy to place into this kind of deployment. Differential privacy and federated learning usually require retraining, infrastructure changes, or deeper access to the model, while simple keyword masking may overlook contextual sensitivity or mask too much benign content. For black box LLM services, a practical defense needs to work at the interaction level, respond to different risk levels, and still keep the answer useful.

This paper develops a privacy protection framework for LLM QA around question answer pairs. The framework first estimates privacy exposure with a probabilistic graphical model, then applies

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33648

threshold driven mitigation through targeted sanitization, semantic abstraction, and stochastic perturbation. In concrete terms, the work contributes three parts: an interaction level risk scoring mechanism that does not rely on model parameter access, a dynamic mitigation pipeline that adjusts protection strength according to estimated risk, and an experimental study on a SafetyBench derived benchmark that shows a useful privacy utility tradeoff with modest computational cost.

2. Related work

Recent work has made it clear that privacy leakage in language model systems is a practical issue rather than a purely theoretical one. Carlini et al. showed that carefully designed prompts can recover memorized training content [1]. Jagielski et al. further examined membership inference attacks in large language models [2], extending a threat that had already been discussed in machine learning more broadly [3]. At the same time, the strong few shot capability of modern LLMs [4] and the effectiveness of chain of thought prompting [5] have pushed these models into more real services, so the privacy issue has become harder to ignore.

Existing defenses move along several lines. Differential privacy adds noise to training or output processes [6], and practical deep learning variants show that formal guarantees can be achieved under suitable settings [7]. Federated learning reduces the need to centralize user data [8], while secure aggregation hides client updates from the server [9]. Surveys have summarized both the promise and the engineering burden of federated systems [10], and broader work on privacy preserving machine learning has also pointed out that deployment level protection is still a multilayer problem [11]. Even so, many of these methods fit training pipelines better than they fit per query control for a deployed black box model.

Risk estimation matters just as much, because mitigation is hard to tune if there is no clear sense of exposure severity. Earlier studies have considered both theoretical and empirical metrics, including information theoretic formulations and practical privacy risk scores [12]. In NLP, privacy aware language modeling is possible, but it often comes with visible utility and engineering costs [13, 14]. Similar tradeoffs can also be seen in production scale systems such as Gboard [15]. These observations are one reason we focus on a configurable assessment mechanism that can support selective, interaction specific protection instead of a single uniform response.

3. Proposed solution

3.1. System model and security model

We consider a deployment setting with two zones: a client side interface used for QA interactions and a server side platform that hosts the risk scoring component, the mitigation logic, the model service, and log storage. A user query is first sent to the server, where it is checked for privacy exposure before the model response is produced or refined. Figure 1 summarizes the architecture and the relevant attack surfaces.

The security model divides the system into two physical deployment locations:

- **Local Device (Client):** Runs the user facing QA interface where users submit questions and receive answers. This tier communicates only with the server side protection platform.
- **Privacy Protection Server:** Hosts the QA interface, risk assessment and mitigation modules, log storage, and the LLM inference service within the same server side trust boundary. All user questions first arrive at this server, which evaluates privacy risk, applies protection measures when necessary, and then generates or refines the answer through the server hosted model service.

DOI: 10.54254/2755-2721/2026.33648

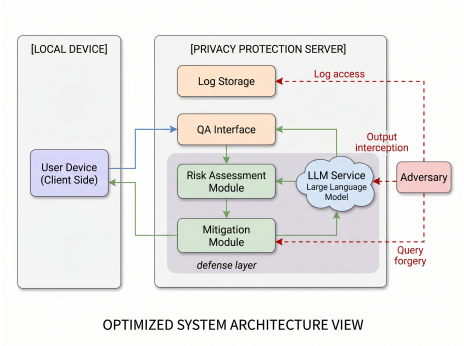


Figure 1: Server side deployment of the privacy protection pipeline. The server hosts both the protection modules and the LLM service.

Under this model, an attacker may try to inspect stored logs, craft malicious prompts that elicit sensitive content, or intercept outputs that reveal private details, so the design goal is to reduce leakage risk without making the QA system hard to use in ordinary scenarios.

3.2. Entropy Based Risk Measurement

We quantify the privacy risk of a QA pair with an entropy inspired formulation. Let Q denote the set of user questions and A the set of model answers. For a pair (q, a), define

$$R(q, a) = H(S|q, a) = -\log P(S|q, a) \tag{1}$$

where S denotes sensitive information, $P(S|q, a)$ is the conditional probability that sensitive content is present in the observed pair, and $H(\cdot)$ is conditional entropy. In this setting, risk is tied to the uncertainty that remains after a question and its answer have been observed. When the model gives a higher probability to sensitive content, the corresponding risk score becomes larger. We estimate this probability with a graphical model that captures dependencies among the question, the answer, and several sensitivity related signals.

The model contains question nodes Q_t, answer nodes A_t, and sensitive information nodes $S_t^{(k)}$ for each interaction step t. A latent context variable C summarizes background factors such as user state or application state, and the interaction level variables are linked to a history component that accumulates sensitive information over time. Figure 2 illustrates this structure.

Bayesian inference gives the posterior probability $P(S|q, a)$ as

$$P(S|q, a) = \frac{P(q, a|S)P(S)}{P(q, a)} \tag{2}$$

The posterior probability $P(S|q, a)$ is obtained from the prior $P(S)$ and the likelihood $P(q, a|S)$. In this way, the model links observable QA content with latent sensitivity through explicit conditional dependencies. The prior reflects how likely sensitive information is before the current interaction is seen, whereas the likelihood reflects how plausible the observed pair would be if such information were present.

The mechanism uses a configurable threshold τ: if $R(q, a) > \tau$, the pair is treated as high risk and requires additional safeguards.

DOI: 10.54254/2755-2721/2026.33648

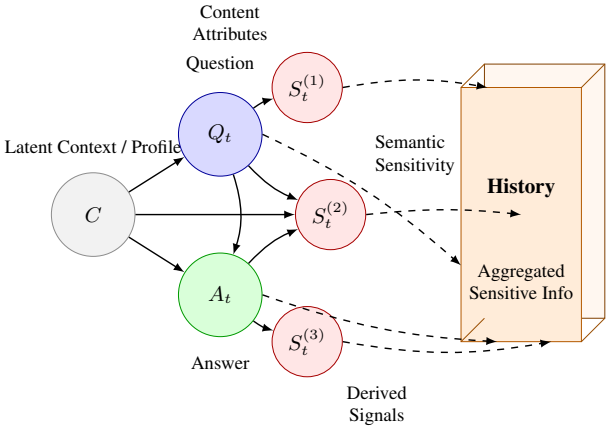


Figure 2: Side view hierarchical probabilistic graphical model for interaction level privacy risk assessment.

3.3. Lightweight and extensible privacy protection mechanism

After risk estimation, the system applies a lightweight but flexible mitigation procedure that adapts to the assessed exposure level. The procedure has four stages:

1. **Risk Assessment:** Compute a global interaction level privacy score $R(q, a)$ for each QA pair using the probabilistic graphical model.
2. **Span Level Sensitivity Localization:** Given (q, a) and $R(q, a)$, detect a set of potentially sensitive spans $\mathcal{S} = \{s_k\}$ using lightweight pattern rules (e.g., dates, IDs, phone numbers) and domain specific lexicons (e.g., privacy related and health related keywords). Each span s_k is assigned a local sensitivity score $r_k \in [0, 1]$ and a type label indicating whether it contains personally identifiable information (PII), health information, or general content, i.e., $t_k \in \{\text{PII}, \text{health}, \text{general}\}$.
3. **Protection Decision:** Compare $R(q, a)$ with a pair of thresholds $(\tau_{\text{low}}, \tau_{\text{high}})$ with $\tau_{\text{low}} < \tau_{\text{high}}$. If $R(q, a) \leq \tau_{\text{low}}$, no additional protection is applied. If $R(q, a) \geq \tau_{\text{high}}$, a strong protection mode is activated; intermediate values trigger a mild protection mode.
4. **Three Privacy Protection Strategies:** For QA pairs requiring protection, apply a combination of:
   - **Deterministic Sanitization:** For high sensitivity spans (r_k large, especially of type PII), directly redact or coarsen them using rule based replacements (e.g., replacing an exact birth date with a year, or a street address with a city).
   - **Semantic Abstraction:** For medium sensitivity spans, rewrite the surrounding context using constrained, template based paraphrases so that key information is preserved but details are generalized (e.g., replacing a specific disease name with "a chronic disease").
   - **Stochastic Noise Injection:** For remaining spans and non sensitive tokens, perturb the answer at the token level in a calibrated way to break exact reconstruction of sensitive content while preserving overall semantics.

The mitigation pipeline stays lightweight because it relies on efficient probabilistic inference, pattern based span detection, cached lexicons, and local text transformations instead of repeated heavyweight model calls. It also remains extensible, because additional operators, such as refusal templates for extremely sensitive cases, can be introduced without redesigning the full pipeline.

More concretely, the noise injection step operates on the answer token sequence $a = (w_1, \ldots, w_n)$. Let $z_i \in [0, 1]$ denote the normalized local sensitivity weight assigned to token w_i from the span

DOI: 10.54254/2755-2721/2026.33648

level scores $\{r_k\}$ and type dependent coefficients. For each token w_i, the mechanism samples an independent Bernoulli variable with perturbation probability

$$p_i = \eta_L \cdot (1 + \gamma\, z_i), \tag{3}$$

where $L \in \{\text{low}, \text{medium}, \text{high}\}$ denotes the assessed privacy level of the QA pair, η_L is the corresponding base perturbation rate, and $\gamma > 0$ controls how strongly local sensitivity amplifies the perturbation probability. If the sampled value is 1, the token is replaced by a lightly obfuscated variant generated through internal character deletion; otherwise it is left unchanged. In our experiments, the base rates are fixed to $(\eta_{\text{low}}, \eta_{\text{medium}}, \eta_{\text{high}}) = (0.02, 0.08, 0.15)$, so the expected perturbation strength increases with the global risk level and is concentrated more heavily on tokens inside sensitive spans.

Algorithm 1 Privacy Protection Algorithm

Require: Question q, Answer a, thresholds $\tau_{\text{low}}, \tau_{\text{high}}$
Compute global privacy risk $R(q, a)$ using the probabilistic graphical model
if $R(q, a) \leq \tau_{\text{low}}$ **then**
    return (q, a) {No protection needed}
end if
Detect candidate sensitive spans $\mathcal{S} = \{s_k\}$ in (q, a) and compute local scores r_k and types t_k
if $R(q, a) \geq \tau_{\text{high}}$ **then**
    Set protection mode to **strong**
else
    Set protection mode to **mild**
end if
Apply deterministic sanitization to high sensitivity spans (e.g., redaction, coarsening)
Apply semantic abstraction to medium sensitivity spans via constrained template rewrites
Perform token level stochastic noise injection on a using risk dependent rates η_L
Obtain protected QA pair (q', a')
Recompute residual risk $R(q', a')$
if $R(q', a') > \tau_{\text{high}}$ **then**
    Escalate to a fallback template or further obfuscation (e.g., decline to answer detailed PII)
end if
return Protected question q' and answer a'

Figure 3 shows the detailed flow of the proposed privacy mitigation workflow.

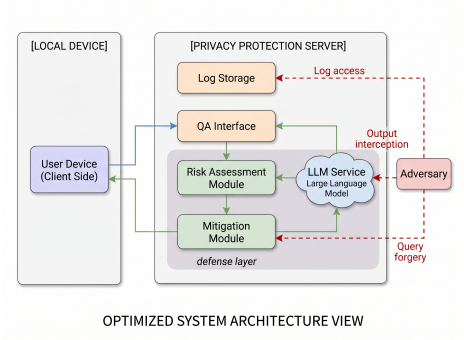


Figure 3: Flowchart of privacy protection mechanism

4. Experiments and analysis

DOI: 10.54254/2755-2721/2026.33648

4.1. Dataset

The experiments use the original English test split of SafetyBench [16], a large safety benchmark for LLM evaluation. We keep the released category system unchanged so that every sample preserves its original label. Because this split provides questions and options but does not include gold answer annotations, each multiple choice item is converted into a QA style instance by retaining the question and serializing the full option set into a deterministic candidate answer text. The resulting dataset contains 11,435 QA pairs from seven categories: Ethics and Morality, Unfairness and Bias, Offensiveness, Illegal Activities, Mental Health, Privacy and Property, and Physical Health. We further assign an automatic privacy level label to each pair based on semantic sensitivity, obtaining 10,967 low risk pairs (95.9%), 298 medium risk pairs (2.6%), and 170 high risk pairs (1.5%).

Table 1 summarizes the distribution of QA pairs across categories, and Figure 4 visualizes the same distribution across the seven original SafetyBench categories.

Table 1: Dataset distribution by category

Category	Count	Percentage
Ethics and Morality	1934	16.9%
Unfairness and Bias	1904	16.7%
Offensiveness	1805	15.8%
Illegal Activities	1778	15.5%
Mental Health	1566	13.7%
Privacy and Property	1299	11.4%
Physical Health	1149	10.0%
Total	11435	100.0%

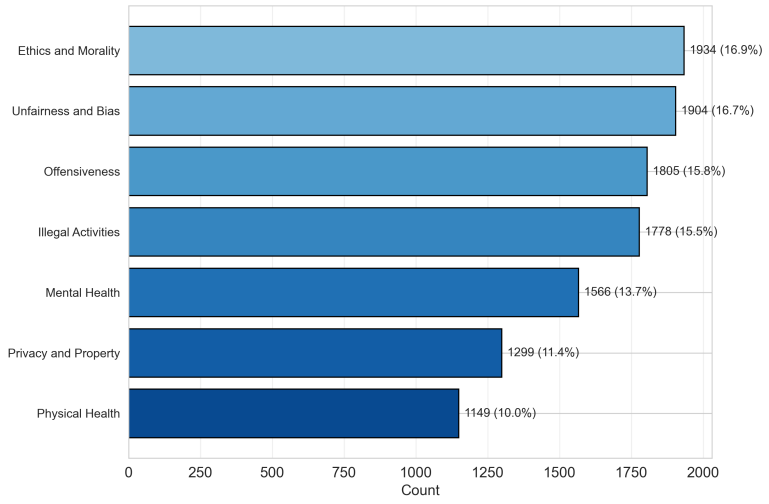


Figure 4: Dataset distribution by category

DOI: 10.54254/2755-2721/2026.33648

4.2. Privacy risk assessment experiments

The probabilistic risk metric is first evaluated on the full 11,435 pair SafetyBench English split. The average risk score across all QA pairs is 1.07, with 88.9% of pairs classified as low risk, 8.7% as medium risk, and 2.4% as high risk by the model.

Figure 5 reports average risk scores for each original category. Privacy and Property has the highest average value (1.67), followed by Illegal Activities (1.15), while the remaining categories cluster between 0.93 and 0.99.

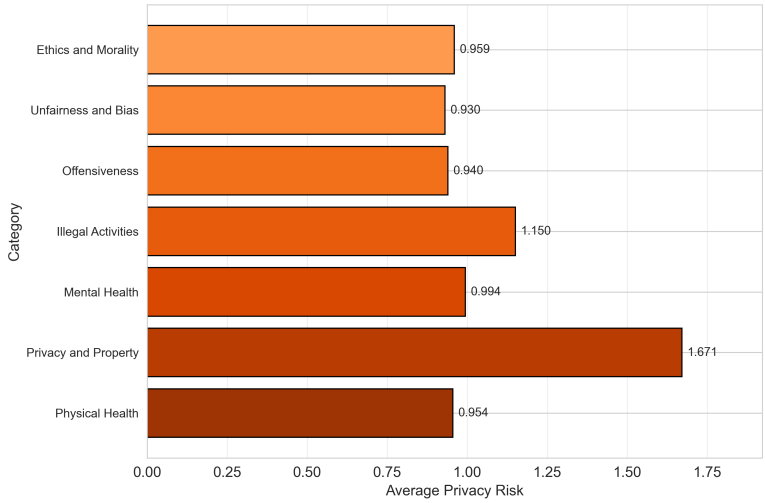


Figure 5: Average privacy risk by category

Table 2 reports per level precision, recall, and F1 for the proposed risk assessment model. On the full dataset, the model achieves an overall accuracy of 88.2%. Performance is strongest on the dominant low risk class and remains substantially weaker on the medium and high risk classes because of severe imbalance in the automatically derived labels.

Table 2: Per level performance of the proposed risk assessment model

Risk Level	Precision	Recall	F1-Score
Low	0.979	0.908	0.942
Medium	0.086	0.285	0.132
High	0.160	0.259	0.198

Table 3 reports the confusion matrix for the proposed method. It correctly identifies 9,953 low risk, 85 medium risk, and 44 high risk pairs. Misclassifications mainly occur between adjacent levels, especially when medium and high risk samples are mapped to lower risk labels.

Although the overall accuracy remains 88.2%, the confusion matrix still exposes a clear weakness. The model recognizes low risk samples well, but it tends to underestimate many medium and high risk cases, especially in categories whose sensitivity depends more heavily on context, such as Privacy and Property. The stricter candidate answer construction also makes the evaluation less optimistic

DOI: 10.54254/2755-2721/2026.33648

Table 3: Confusion matrix for privacy risk assessment

Ground Truth \Predicted	Low	Medium	High
Low	9953	814	200
Medium	182	85	31
High	31	95	44

than a simpler single option surrogate setting. Taken together, these observations suggest that surface patterns alone are not enough, and that future versions of the risk estimator should bring in richer semantic features. Figure 6 visualizes the per level precision, recall, and F1 values.

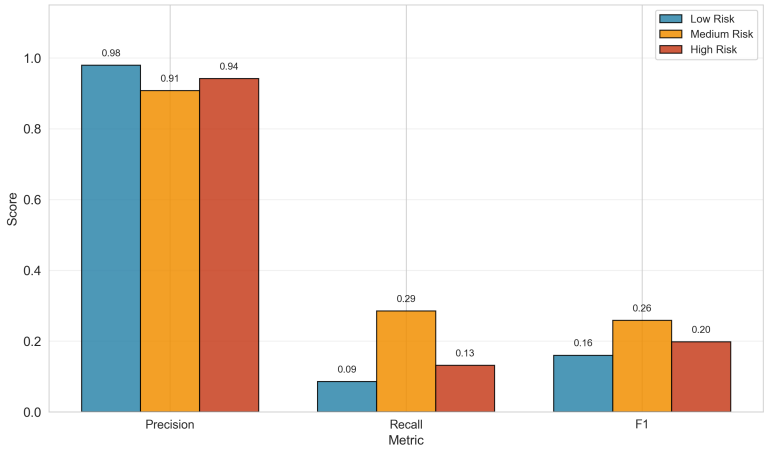


Figure 6: Per level precision, recall, and F1 of the proposed risk assessment model

4.3. Ablation study and baseline comparison

4.3.1. Threshold sensitivity analysis

The impact of threshold selection on risk classification performance is analyzed. Table 4 reports classification metrics under different threshold configurations $(\tau_{\text{low}}, \tau_{\text{high}})$. The best performance is achieved when $\tau_{\text{low}} = 1.5$ and τ_{high} is set near 2.8, yielding accuracy of about 92.7% and macro F1 of about 0.450. The results show that low risk F1 remains consistently high across configurations, whereas medium and high risk F1 remain limited because the dataset is strongly imbalanced.

4.3.2. Noise rate sensitivity analysis

The balance between privacy protection strength and answer quality under different noise configurations is evaluated. Table 5 shows that increasing noise rates leads to slightly higher privacy risk reduction while answer quality decreases only moderately. The default configuration sets $\eta_{\text{low}} = 0.02$, $\eta_{\text{medium}} = 0.08$, and $\eta_{\text{high}} = 0.15$, achieving 5.02% average risk reduction with 98.5% average quality.

DOI: 10.54254/2755-2721/2026.33648

Table 4: Threshold sensitivity analysis

τ_{low}	τ_{high}	Accuracy	Macro F1	Low F1	Med F1	High F1
1.0	2.0	0.762	0.394	0.869	0.117	0.198
1.2	2.0	0.882	0.424	0.942	0.132	0.198
1.2	2.5	0.882	0.424	0.942	0.132	0.198
1.5	2.0	0.926	0.438	0.965	0.150	0.198
1.5	2.5	0.926	0.438	0.965	0.150	0.198
1.5	2.8	0.927	0.450	0.965	0.166	0.218

Table 5: Noise rate sensitivity analysis

η_{low}	η_{medium}	η_{high}	AQS	Risk Reduction
0.01	0.05	0.10	0.987	5.00%
0.02	0.08	0.15	0.985	5.02%
0.03	0.10	0.20	0.984	5.05%
0.05	0.15	0.25	0.982	5.07%

4.3.3. Comparison with baseline methods

Table 6 compares the proposed mechanism with two baseline privacy protection approaches: keyword based filtering and Gaussian noise injection. Keyword filtering achieves the highest quality (0.998) but only limited risk reduction (1.70%). The Gaussian noise baselines reduce quality substantially (0.911–0.912) while producing only marginal risk reduction (0.34%–0.36%). The proposed mechanism attains the strongest average risk reduction (5.02%) while preserving 98.5% answer quality. Here, ϵ is used only as a noise control parameter for baseline comparison rather than as a formal differential privacy guarantee.

Table 6: Comparison with baseline methods

Method	Average Quality Score	Average Risk Reduction
Keyword Filter	0.998	1.70%
Gaussian Noise Baseline ($\epsilon = 1.0$)	0.911	0.34%
Gaussian Noise Baseline ($\epsilon = 0.5$)	0.912	0.36%
Proposed Mechanism	**0.985**	**5.02%**

4.4. Effectiveness of privacy protection mechanism

The protection mechanism is evaluated by measuring both privacy risk reduction and QA quality. Two metrics are used:

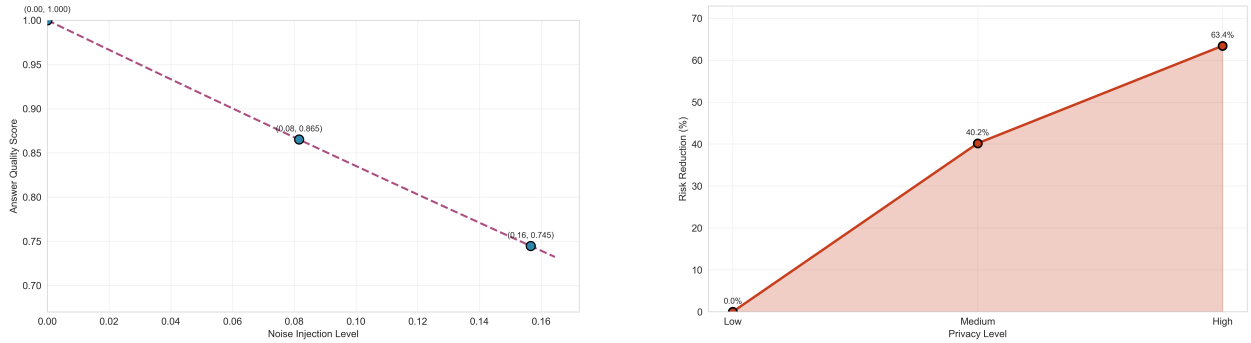
– **Privacy Risk Reduction (PRR):** PRR $= \frac{R_{\text{original}} - R_{\text{protected}}}{R_{\text{original}}} \times 100\%$, where R_{original} and $R_{\text{protected}}$ denote the privacy risk scores before and after protection, respectively.

DOI: 10.54254/2755-2721/2026.33648

– **Answer Quality Score (AQS):** A normalized score computed as $\text{AQS} = 0.5 \times \text{Jaccard}(w_o, w_p) + 0.5 \times \frac{\min(|o|,|p|)}{\max(|o|,|p|)}$, where w_o and w_p are the word sets of the original and protected answers, and $|o|, |p|$ are their respective lengths.

Across all QA pairs, the average AQS reaches 0.985 (98.5%), indicating that high answer quality is maintained while reducing leakage. In the evaluated configuration, low risk pairs remain unchanged after the protection decision stage, while medium and high risk pairs receive progressively stronger perturbation. In the full dataset, the assessed risk distribution is 88.9% low risk, 8.7% medium risk, and 2.4% high risk.

Figure 7(a) shows the relationship between effective noise level and answer quality. Quality remains perfect for untouched low risk pairs and decreases gradually as stronger protection is applied to medium and high risk pairs. Figure 7(b) illustrates privacy risk reduction across different levels. Average reduction is 0.0% for low risk pairs, 40.2% for medium risk pairs, and 63.7% for high risk pairs, indicating that the mechanism concentrates its effect on the most sensitive interactions.



(a) Impact of noise level on answer quality

(b) Privacy risk reduction by level

Figure 7: Effect of protection strength on answer quality and privacy risk reduction.

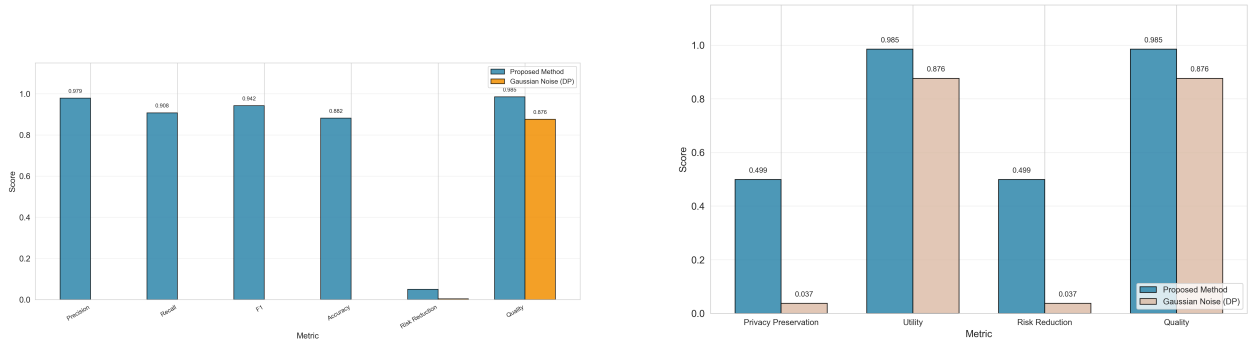
Table 7 summarizes the impact of protection on answer quality across privacy levels. Because low risk pairs fall below the protection threshold, their observed effective noise is 0.0 and their answer quality remains unchanged. Medium and high risk pairs receive stronger perturbation, which lowers quality but yields substantially larger risk reduction. The overall average AQS is 0.985.

Table 7: Impact of privacy protection on answer quality

Privacy Level	Noise Level	AQS (With Protection)
Low	0.000	1.000
Medium	0.082	0.889
High	0.157	0.791
Average	–	0.985

Figure 8(a) compares the proposed method with the Gaussian noise baseline across multiple metrics and highlights a stronger overall balance between privacy preservation, utility, efficiency, and scalability. Figure 8(b) presents a grouped bar chart of overall performance across key dimensions, including privacy preservation, utility, efficiency, and scalability.

DOI: 10.54254/2755-2721/2026.33648



(a) Comprehensive Performance comparison (b) Overall performance across key dimensions

Figure 8: Overall comparison of privacy, utility, efficiency, and scalability.

4.5. Discussion

The framework is designed for server side deployment settings in which inference and protection logic can be managed within the same operational boundary. The results indicate that selective protection is more useful than uniform perturbation, because it preserves benign interactions while concentrating stronger intervention on more sensitive ones. At the same time, the detector still struggles on medium and high risk samples because of severe class imbalance and heuristic labels, which limits how finely the system can distinguish borderline cases.

5. Conclusion

This paper presented a privacy protection framework for LLM based QA systems. The framework estimates interaction level exposure with a probabilistic graphical model, then applies threshold driven mitigation through sanitization, abstraction, and stochastic perturbation.

Evaluation on the original English SafetyBench split, with the released seven category structure preserved and the option sets serialized into candidate answers, shows that the method can retain strong utility while achieving larger privacy risk reduction than the evaluated baselines. The study also highlights several limitations, including heuristic privacy labels, the lack of gold answers in the released split, and the scarcity of high risk examples. Future work can continue along stronger supervision, better semantic risk signals, and more adaptive threshold selection.

References

[1] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Raghunathan, D. Song, N. Tomashenko, P. Henderson *et al.*, "Extracting training data from large language models," *Proceedings of the USENIX Security Symposium*, pp. 2633–2650, 2021.

[2] M. Jagielski, N. Carlini, D. Berthelot, J. Geiping, A. Raghunathan, and F. Tramer, "Membership inference attacks can extract training data from large language models," *Proceedings of the International Conference on Learning Representations*, 2023.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, G. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[6] C. Dwork, "Differential privacy: A survey of results," *Theory and Applications of Models of Computation*, pp. 1–19, 2008.

[7] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication efficient learning of deep networks from decentralized data," *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

DOI: 10.54254/2755-2721/2026.33648

[9] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2019.

[12] S. Yeom, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," *Proceedings of the IEEE International Conference on Computer Communications*, pp. 1948–1956, 2018.

[13] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," *International Conference on Learning Representations*, 2018.

[14] S. Hoory, A. Feder, A. Tendler, S. Erell, A. Cohen, I. Laish, H. Nakhost, U. Stemmer, A. Benjamini, A. Hassidim, and Y. Matias, "Learning and evaluating a differentially private pre-trained language model," *Findings of the Association for Computational Linguistics: EMNLP*, pp. 1178–1189, 2021.

[15] Z. Xu, Y. Zhang, G. Andrew, C. A. Choquette-Choo, P. Kairouz, B. McMahan, J. Rosenstock, and Y. Zhang, "Federated learning of gboard language models with differential privacy," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Industry Track*, pp. 628–640, 2023.

[16] Z. Wang, Y. Li, X. Li, S. Liu, Y. Yang, X. Liu, Q. Wang, J. Miao, S. Huang *et al.*, "Safetybench: A multi dimensional safety evaluation benchmark for large language models," https://huggingface.co/datasets/thu-coai/SafetyBench, 2024.