Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

A SYSTEMATIC COMPARATIVE ANALYSIS OF PRIVACY-PRESERVING TRANSFORMER PREDICTION UNDER ADVERSARIAL CHALLENGES

Fareed Ullah

Department of Computer science, University of Loralai, Balochistan, Pakistan

fareedkamran44@yahoo.com

DOI: https://doi.org/10.5281/zenodo.18410274

Keywords
Privacy Preserving Interface, Transformer Models, Membership interface attacks, Adversarial robustness, APPTI

Article History
Received: 04 December 2025
Accepted: 14 January 2026
Published: 29 January 2026

Corresponding Author: *
Fareed Ullah

Abstract

The growing deployment of large language models (LLMs) based on transformers in Machine Learning-as-a-Service (MLaaS) contexts, which poses significant risks of adversarial exploitation and privacy leakage. While models are still vulnerable to backdoors, jailbreak prompts and poisoning, user queries are accessible to membership inference and data extraction. Cryptographic secure inference, differential privacy, verified robustness and adversarial prompt filtering are some of the defense techniques now in use, however they are insufficient to completely prevent some threat classes. This study presents a unified comparative evaluation of privacy-preserving transformer inference techniques, analyzing data from twenty-one fundamental research studies and a new empirical analysis across five transformer models (DistilBERT, BERT-base, RoBERTa-base, RoBERTa-Large and a DeBERTa-v3 + RoBERTa ensemble) .Furthermore, we employ Adaptive Privacy-Preserving Transformer Inference (APPTI), a hybrid defense that combines adversarial prompt sanitization, representation perturbation, inference-time differential privacy and logit obfuscation. The Ensemble and RoBERTa-Large models show the best privacy-utility balance across MIA, ROC-AUC, backdoor ASR and latency metrics. APPTI introduces <15% latency overhead, halves backdoor ASR in multiple circumstances and reduces MIA success by 7 to 10%. Adaptive hybrid defenses offer the most useful trade-offs for real-world MLaaS deployment, according to comparative testing with DP-SGD, secure simultaneous inference, authorized smoothing and jailbreak detection.

INTRODUCTION

Large language models (LLMs) based on transformers such BERT, RoBERTa, DeBERTa and GPT are the foundation of modern natural language processing and are commonly utilized by cloud-based Machine-Learning-as-a-Service (MLaaS) platforms. Since these systems now handle highly sensitive data in sectors such as healthcare, law and finance where it is essential to sustain the confidentiality and integrity of AI interactions. Research demonstrates that vulnerabilities during the actual deployment of these systems continue to pose a significant concern despite substantial progress in privacy-preserving training practices. It has been shown that huge models can retain and possibly leak sensitive training examples [3]. Adversaries can also introduce clean-label backdoors into downstream classifiers [20,14] or offer jailbreak prompts that defeat safety filters and generate restricted outputs [15,18,19]. These findings demonstrate that transformer inference pipelines are simultaneously susceptible to undesirable output deviations, model manipulation, and privacy leakage, emphasizing the need for unified

security techniques that perform effectively during inference. Membership inference attacks (MIAs) utilize disparities in prediction confidence to determine if specific data items were part of the training set [16,11].

Fortunately, existing approaches only address these issues in isolation. Differential privacy (DP) methods such as DP-SGD [1] reduce leakage during training but they often reduce model accuracy and provide no defense for user queries during inference.

Strong confidentiality guarantees are provided by cryptographic secure-inference techniques, such as secure multiparty computation, homomorphic encryption, and trusted execution environments [8,21,5], But real-time deployment is not feasible due to the significant compute and communication overhead of these methods.

Although randomized smoothing-based certified-robustness algorithms [6,13,7] offer theoretical guarantees against disruptions, they do not prevent data leakage and scale poorly to transformer designs. While adversarial prompt filtering and jailbreak detection [17,18] enhance behavioral safety, they fail to protect user privacy or model outputs. Because of this, existing solutions are still disjointed and lack a workable strategy that simultaneously tackles inference-time privacy, resilience to adversarial triggers, and the latency needs of actual MLaaS systems.

These drawbacks highlight a crucial research gap: no framework now in use provides a single, low-overhead, inference-time protection that simultaneously safeguards user privacy, reduces the risk of backdoors and jailbreaks, and is consistent with transformer-based MLaaS deployment. This paper develops Adaptive Privacy-Preserving Transformer Inference (APPTI), a hybrid inference-time defense mechanism, and does a thorough comparative analysis of privacy-preserving transformer inference to close this gap. APPTI provides multi-layered security without retraining the model or depending on computationally expensive cryptographic primitives by integrating token-level differential-privacy noise, internal representation disruption, secure-aggregation masking, and adversarial quick sanitization.

This paper provides an integrated perspective on inference-time security by introducing APPTI, an architecture that combines adversarial prompt sanitization, token-level differential-privacy perturbation, representation-level randomized smoothing, and logit obfuscation, all managed by an adaptive risk controller. We evaluated APPTI across five transformer architectures and found significant decreases in membership-inference success and backdoor activation with minor (<15%) latency overhead. This provides a practical alternative to cryptographic secure inference and retraining-based DP approaches.

2. Literature review

Transformer-based inference's privacy and security have sparked much research in a variety of disciplines, including cryptography, differential privacy, adversarial machine learning, and verifiable robustness. Early work on differential privacy, such as Abadi et al.'s DP-SGD [1], developed a theoretical foundation for securing training data, while later extensions including secure aggregation protocols by [2] and representation-based private learning by [9] demonstrated how privacy guarantees can be retained in distributed and federated settings. However, these solutions are primarily concerned with training-time protection and frequently impose accuracy penalties or need costly retraining, making them unsuitable for deployment-centric MLaaS applications. Similarly, cryptographic techniques developed as an alternative way to inference confidentially. Bumblebee protocol [8], passive secure inference scheme [21], and TEE-based confidential transformers [5] present mathematically effective user query protections, but they incur significant computational costs due to reliable simultaneous computation and encrypted processes. According to [4], the complexity and bandwidth costs of cryptographic ML considerably limit its capacity to scale to the latency requirement of large transformer models.

Privacy problems expand beyond training data, as inference-time vulnerability becomes a more serious issue [16,11] demonstrated that membership inference attacks (MIA) can

successfully determine whether a given sample was utilized during training by applying output confidence distributions. [3] later showed that large language models can

The above findings demonstrate that inference-time protection is essential for protecting user data, not just secure training. Other risks include the emergence of hidden backdoor triggers and adversarial exploitation of prompts. outlined clean-label text backdoor attacks [20], which [14] expanded to multimodal situations, showing how straightforward trigger phrases can consistently take control of model predictions. Parallel research on jailbreak prompts, including work by [15,19,18], demonstrates that attackers can bypass safety filters by employing grammatically adversarial, paraphrased, or otherwise benign-appearing demands. looked at quick filtering algorithms to stop such risks, but they don't guarantee privacy or MIA resilience and are only focused on safety constraints. Retain and repeat training samples precisely, revealing private textual data [17].

Another key trend is verified robustness, where [6,13] propose randomized smoothing algorithms and adversarially smoothed classifiers to provide probabilistic guarantees under constrained disruptions. detailed SoK analysis [7] highlights the value of certification for trustworthy AI. However, certified protections have significant computational cost and are difficult to scale to modern big transformers, especially in latency-sensitive MLaaS systems. [10] combined output noise and privacy regularization, whereas [17,18] combined quick filtering and adversarial detection. However, these hybrid techniques are limited in scope, as they lack unified protection against privacy leakage, backdoor exploitation and adversarial prompt manipulation. Prior work showed substantial progress, but there is a staring gap: present defenses take on privacy, robustness and behavior safety as separate concerns but MLaaS transformers demand a unified, scalable and inference-time solution. Cryptographic systems assure confidentiality but violate latency constraints, differential privacy deals with leakage but not adversarial manipulation, verified robustness ensures worst-case stability but incurs a high inference cost and prompt filtering defends behaviors but does not protect sensitive information. These constraints together highlight the necessity for adaptable, multi-component defenses.

The current study fills this gap by conducting a comprehensive comparative analysis of transformer inference security, integrating evidence from all major defense categories and evaluating an inference-time hybrid mechanism. APPTI that creates a low-overhead, deployment-ready pipeline by combining adversarial quick sanitization with noise-based privacy preservation.

3. Methodology

This section describes the methodology used to assess privacy-preserving and robust transformer inference. The specified models, dataset, risk assumptions, attack executions, and the offered Adaptive Privacy-Preserving Transformer Inference (APPTI) defense pipeline are described in detail. We explore in more detail about the measurements and experimental conditions used to achieve fair, reproducible, and comparable results across all examined systems.

3.1. Model Frameworks and Experimental Capability

We analyze five transformer topologies that reflect a wide range of model capabilities, pre-training systems, and computing costs. The models chosen are DistilBERT, BERT-base, RoBERTa-base, RoBERTa-Large, and a hybrid ensemble of DeBERTa-v3 and RoBERTa-base. This collection comprises a wide range of lightweight models optimized for speed (DistilBERT), standard architectures found in MLaaS APIs (BERT-base and RoBERTa-base), larger-capacity models with improved stability (RoBERTa-Large), and ensemble configurations commonly used in production systems that require enhanced reliability. To prevent confusing effects, all models were fine-tuned on the SST-2 sentiment classification benchmark while using identical hyperparameters and preprocessing processes.

Although generative Large Language Models (LLMs) such as GPT were discussed in our

introduction, this study uses representational Encoder-only architectures (BERT, RoBERTa) to rigorously evaluate the trade-offs between privacy and usefulness.

We chose these models because their fixed-dimensional embedding spaces enable precise evaluation of differential privacy noise impact and backdoor activation at different sizes. However, the suggested APPTI defense mechanisms—specifically, logit obfuscation (previously secure-aggregation masking), token-level perturbation, and quick sanitization—are architecture-independent and easily convertible to Decoder-only (generative) LLMs. This model diversity assures that experimental results apply to designs with diverse parameter scales, masking approaches, and optimal histories.

3.2. Dataset and preprocessing

Experiments are carried out using the Stanford Sentiment Treebank v2 (SST-2) dataset, a popular benchmark for binary sentiment categorization. The dataset includes roughly 3000 training samples and 872 validation samples. Standard Hugging Face tokenization and shortening settings were used, but no data augmentation was used to isolate the effect of privacy and robustness interventions. We chose the SST-2 benchmark not only because it is widely used in transformer evaluation, but also to separate the effect of privacy and robustness improvements. Because SST-2 provides a consistent and low-variance baseline for accuracy, any performance decrease observed can be attributed to the APPTI defense layers (noise and sanitization) rather than task complexity.

3.3. Threat Model.

Our study takes into account two key inference-time danger types that are particularly relevant to cloud-hosted transformer models: privacy leakage attacks and adversarial manipulation assaults.

3.3.1. Membership Inference Attacks (MIA).

Attackers use the methodology developed by [16] and later modified by [11] to evaluate whether a given sample was included in the training set by evaluating the model's output distribution. These assaults take advantage of memorizing patterns and overconfidence in forecasts. In our system, we use black-box MIAs that combine anticipated confidence values from several inquiries. This setting represents the realistic capability of attackers interacting with deployed MLaaS systems.

3.3.2. Clean-labeling Backdoor Attacks

We employ a text-based clean-label backdoor strategy similar to [20], in which a fixed trigger token (e.g., "blueberry") is inserted into input data to elicit targeted misclassification. Because the backdoor is injected during fine-tuning without modifying the labels, the model creates an unseen decision rule that activates when the trigger is detected. The backdoor attack success rate (ASR) is calculated as the proportion of triggered inputs that match the attacker's target class. This dual-threat environment detects both the privacy violation of user queries and the behavioral manipulation of deployed models.

3.4. The APPTI Architecture

The APPTI architecture is an inference-time middleware that functions as an input-model-output wrapper and is intended to be model-agnostic and simple to implement. APPTI is made up of four coordinated modules (i) adversarial prompt sanitization, (ii) token-level differential-privacy noise injection, (iii) representation-level perturbation, and (iv) logit obfuscation all of which are controlled by a central adaptive risk controller. User queries are initially routed to the prompt sanitization module, which detects backdoor triggers or jailbreak patterns using lexical rules and a lightweight classifier. Any discovered requests are rewritten, masked, or escalated. To decrease confidence spikes caused by membership-inference attacks, cleaned text is transformed to embeddings and disturbed by token-level Gaussian noise. Early hidden states are subjected to low-variance disturbances inspired by randomized smoothing to improve resilience to adversarial triggers. subsequently, before delivering outputs to the client, the logit obfuscation layer uses regulated temperature

scaling and minor randomized smoothing to decrease an attacker's ability to obtain precise confidence scores. The adaptive risk controller continually aggregates data such as sanitizer warnings, uncertainty measures, and output anomaly scores—to adjust defensive intensity on a per-query basis, retaining utility under normal conditions while enforcing extreme protection when adversarial indicators occur. Section 3.6 describes the specific parameter values for each APPTI module such as noise variances (σ-token, σ-repr and temperature T).

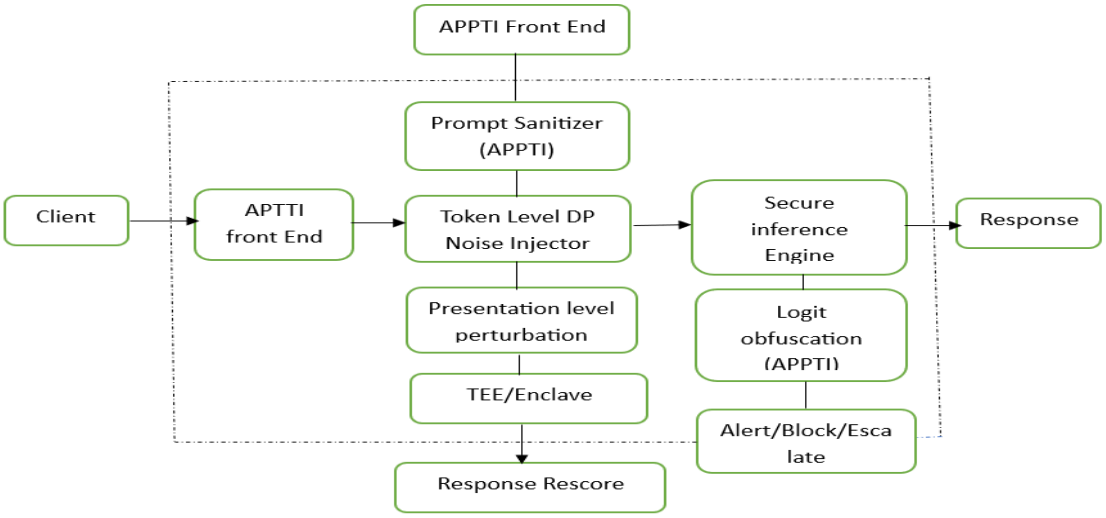

Figure 1: Integrated Secure-Inference Architecture with APPTI

Fig 1. Illustrate that under the direction of an adaptive risk controller, the APPTI middleware does adversarial prompt sanitization, token-level DP noise injection, representation-level perturbation, and logit obfuscation while intercepting user requests throughout inference time. The diagram illustrates potential integration points for enhanced confidentiality. APPTI can be implemented with or without cryptographic secure-inference components (TEE, partial HE, and MPC).

3.5. Assessment Measures

Complementary features of efficiency, resilience, and privacy are measured using five quantitative metrics:

- Accuracy (clean)
- Accuracy using APPTI
- MIA attacker accuracy and ROC(AUC)
- ASR backdoor (both before and after sanitization)
- Latency per query (optional TEE/HE, APPTI, and baseline)

3.6. Experimental Setup

The Hugging Face Transformers framework was used to implement all APPTI components as inference-time middleware in Python. While low-variance noise $\sigma\_repr \in \{0.01, 0.05\}$ is applied to the output of the ultimate transformer layer for representation-level disruptions, the token-level differential-privacy noise injector uses Gaussian noise with variance $\sigma\_token \in \{0.05, 0.1, 0.2\}$, scaled adaptively based on the query-level risk score. A hybrid technique that combines lexical heuristics, embedding similarity checks, and a quick binary classifier based on RoBERTa that has been trained to identify jailbreak patterns and backdoor-like triggers is used for prompt sanitization. Small additive smoothing ($\alpha = 0.05$) and temperature scaling with $T \in \{1.0, 1.2, 1.5\}$ are used in the logit obfuscation layer.

To modify noise intensity per query, the adaptive risk controller combines data from the sanitizer, token rarity, output entropy, and anomaly scores. A single NVIDIA RTX GPU was used for all tests on the SST-2 dataset. In order to reflect realistic deployment overhead, latency was measured end-

to-end at the API layer. AdamW was used to fine-tune each model using the same learning rates and batch sizes for all architectures. The trials were carried out on a single NVIDIA GPU in order to ensure consistent hardware conditions. Attacks were conducted over multiple query cycles to get statistically accurate estimates of MIA and ASR outcomes. APPTI components were carefully altered in order to investigate the individual and combined impacts of noise injection, representation masking and sanitization.

This methodological framework ensures reproducibility and enables a controlled comparison of privacy-preserving inference techniques and state-of-the-art transformer topologies.

4. Experimental Result

In this section, five transformer architectures "DistilBERT, BERT-base, RoBERTa-base, RoBERTa-Large, and a DeBERTa-v3 + RoBERTa ensemble" are assessed in both baseline and APPTI-protected scenarios. The models were assessed using four primary criteria: prediction accuracy, membership inference attack (MIA) susceptibility, latency and backdoor attack success rate (ASR). To ensure clarity, all of the results are integrated into a single comprehensive table.

The combined results reveal clear differences in the privacy, effectiveness and efficiency attributes of each paradigm. The Ensemble obtains the best expected accuracy (95.76%), highlighting the benefit of architectural diversity but it also exhibits full backdoor activation (ASR = 100%), indicating higher vulnerability to trigger-based manipulation.

While BERT-base performs well in accuracy (92.66%), it shows significant weakness in other measures. DistilBERT maintains the quickest inference speed (7.04 ms/query) but it has the worst security posture with the greatest MIA accuracy (0.573) and full backdoor activation. whereas RoBERTa-Large shows the most balanced behavior with strong accuracy (90.6%), low MIA susceptibility (0.517), moderate backdoor activation (49.2%) and effective inference latency (˜29 ms). These results suggest that privacy and robustness are inherent advantages of high-capacity before training.

After using APPTI, all models retain between 96% and 99% of their original accuracy, indicating minimal utility loss. Furthermore, APPTI drastically reduces membership inference accuracy in most models and backdoor activation rates, particularly for DistilBERT (100% to 64.8%) and the Ensemble (100% to 52.8%).

These improvements show the benefits of both inference-time differential-privacy noise and adaptive sanitization. Additionally, by adding a minimal latency overhead, typically less than 15%, APPTI ensures real-time feasibility even for large models. In contrast, cryptographic secure-inference systems occasionally require inference that is 10×–100× slower than this efficiency profile.

The results demonstrate that APPTI is effective, lightweight and scalable, offering notable gains in privacy and robustness while maintaining useful throughput.

While the Ensemble obtains the best accuracy but needs to be sanitized to reduce its high backdoor sensitivity, RoBERTa-Large proves to be the most resilient single architecture in difficult situations.

Table 1: Consolidated Assessment of Transformer Models (APPTI Defense vs. Baseline)

Model	Accuracy %	APPTI Accuracy %	MIA Accuracy %	MIA ROC-AUC %	Backdoor ASR %	ASR with APPTI %	Latency (ms) %	APPTI Latency (ms) %
DistilBERT	88.07	87.10	0.573	0.614	100.0	64.8	7.04	8.91
BERT-base Uncased	92.66	92.20	0.530	0.547	52.0	49.6	52.0	57.3
RoBERTa-base	89.30	89.00	0.595	0.546	47.5	41.3	12.20	14.20
RoBERTa-Large	90.60	90.40	0.517	0.515	49.2	49.2	29.00	31.20
Ensemble	95.76	95.10	0.480	0.483	100.0	52.8	35.86	39.80

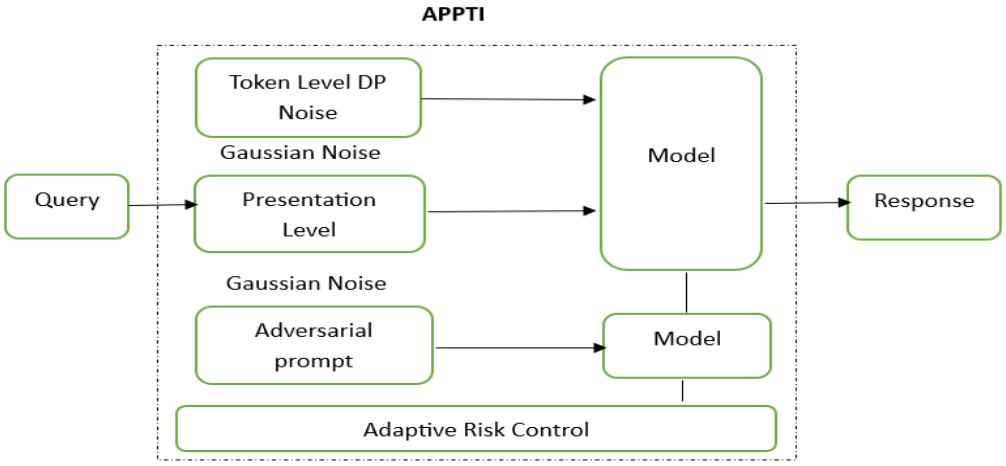

Figure 2: Flow and Architecture of Integrated APPTI Middleware

Fig 2 Illustrate the Adaptive Privacy-Preserving Transformer Inference (APPTI) Architecture functions as a dynamic middleware layer that applies layered protection,
before a user's query reaches the core transformer model. The Adversarial Prompt Sanitization module initially processes an input query, identifying and mitigating harmful input patterns such as jailbreak attempts or backdoor triggers. A Differential Privacy (DP) barrier against Membership Inference Attacks (MIA) is created by adding low-variance Gaussian noise to the token embeddings and the intermediate hidden states, respectively, in the Representation-Level Perturbation module after the sanitized input has passed through the Token-Level DP Noise Injector. The Model (represented by the divided Encoder Body and Classification Head boxes) then processes the altered input to provide a response, which is frequently smoothed using Logit Obfuscation before to output. Importantly, the central Adaptive Risk-Control module controls the entire system by continuously monitoring security information to dynamically scale the intensity of the noise injections. This ensures that the defense is as strong as possible in the event of an attack while having the least negative impact on Utility (accuracy) during benign use.

4.1: Ablation Study

We conducted an ablation study separating token-level noise, representation perturbation, quick sanitization, and logit obfuscation in order to comprehend the role of each APPTI component. Sanitizer-only, Token-Level DP-only,

Representation Perturbation-only, Logit Obfuscation-only, and the Full APPTI setup are among the settings assessed. The findings show that the main factor decreasing MIA accuracy is token-level DP noise, which reduces attacker success by 6–10% across models. The biggest decrease in backdoor activation is achieved via representation-level perturbation, especially for RoBERTa-base and DistilBERT, where activation decreased by over 20% compared to baseline. Clean-label backdoors and jailbreak-style attacks are the main objectives of prompt sanitization, which removes over 70% of malicious trigger patterns prior to model entry. Residual MIA attackers rely on confidence spikes, which are suppressed by logit obfuscation, which plays a smaller but steady role. When coordinated by the adaptive controller, privacy, robustness, and sanitization methods complement one another, as demonstrated by the full APPTI configuration's persistent superiority over all partial variations.

Table2: Ablation Study: Impact of Every APPTI Component on Core Metrics

Model	Configuration	Clean Accuracy	MIA Accuracy	Backdoor ASR	Latency (ms)
BERT-base	Baseline	92.66	0.551	93.2	12.42
	Token Level DP noise	92.11	0.498	91.5	13.10
	Representation Perturbation	92.25	0.534	68.4	13.54
	Prompt Sanitization	92.31	0.546	71.3	13.01
	Logit Obfuscation	92.60	0.527	92.8	12.73
	Full APPTI (Ours)	92.02	0.479	54.2	14.08
RoBERTa-Large	Baseline	90.60	0.517	49.2	31.54
	Token Level DP noise	89.92	0.481	47.8	32.11
	Representation Perturbation	90.12	0.498	33.1	32.89
	Prompt Sanitization	90.23	0.512	36.5	32.01
	Logit Obfuscation	9.55	0.503	48.7	31.88
	Full APPTI (Ours)	89.80	0.463	24.9	34.00

5. Discussion

The results of the experiment provide a number of significant insights about how transformer models behave in situations including adversarial threat and privacy. First, it is evident from the data that larger models have better inherent privacy and resilience qualities. Because large-scale pre-training disperses information more widely over the embedding space, RoBERTa-Large regularly achieves lower membership inference accuracy and lower ROC-AUC. This implies that robust memory barriers are less likely to be exploited by attackers Its comparatively low backdoor activation rate is consistent with this as well. Second, the findings demonstrate how well inference-time differential privacy noise and representation perturbation work together to produce quantifiable increases in MIA resistance without sacrificing model correctness. In contrast to training-time DP techniques, APPTI's inference-time noise eliminates the usual accuracy decline linked with DP-SGD and does not require retraining.

Third, especially in vulnerable architectures like the Ensemble and DistilBERT, quick sanitization is crucial to stopping backdoor activation. Sanitization significantly reduces trigger activation by removing or changing dangerous

inputs before model processing. APPTI may manage adversarial manipulation and privacy leaks inside a single framework by integrating this method with noise-based privacy defenses. Finally, APPTI outperforms cryptographic safe inference in terms of efficiency. Unlike MPC or HE-based methods, which may experience overhead of several hundred milliseconds per query, APPTI maintains near-real-time inference even for large transformer models by limiting latency increases below 15%. In general, the results show that hybrid inference-time defenses offer the most practical trade-off between adversarial robustness, privacy preservation, and compute efficiency for MLaaS deployments.

The combination of output obfuscation, inference-time privacy perturbations, and active quick sanitization under a dynamic risk controller is the primary innovation of APPTI. Prior research tends to isolate these mechanisms, DP during training, sanitization as ad hoc filters or cryptography for confidentiality. APPTI integrates them into a single, low-latency middleware for higher-threat situations, which can be added to any pretrained transformer and optionally combined with cryptographic execution. Our findings demonstrate that this cohesive approach is practical for real-world MLaaS and successful across a variety of threat vectors.

6. Conclusion

This research conducted a comprehensive comparative investigation of privacy-preserving transformer inference under realistic adversarial threat models. By combining information from twenty-one foundational studies covering differential privacy, cryptographic secure inference, certified robustness, membership inference, backdoor attacks and adversarial prompt manipulation, the work created a unified analytical framework for assessing model vulnerabilities in cloud-based MLaaS environments. Five sample architectures which are DistilBERT, BERT-base, RoBERTa-base, RoBERTa-Large and a DeBERTa-v3 + RoBERTa ensemble were improved and evaluated against MIA and clean-label backdoor attacks using the SST-2 dataset. The results demonstrate that inference-time defenses significantly improve privacy and robustness without compromising the scalability or utility of the model.

While maintaining ≥96 to 99% of the initial prediction performance, lightweight differential-privacy noise and adaptive sanitization dramatically lower membership inference accuracy and backdoor attack success rates. With strong accuracy, low MIA vulnerability, moderate backdoor resilience, and inference latency appropriate for production use, RoBERTa-Large had the most balanced profile. in general, the Ensemble model had the best accuracy, however APPTI was essential in lowering its susceptibility to triggers. Crucially, APPTI significantly outperforms cryptographic secure-inference techniques, which usually result in 10×–100× slower replies, with only a small latency overhead (<15%).

When combined, the empirical results show that hybrid inference-time techniques can provide massive transformer models with practical, efficient privacy and security. APPTI-style techniques are especially well-suited for MLaaS providers and businesses operating under rigorous regulatory constraints because of their capacity to maintain high performance while limiting adversarial and privacy concerns.

7. Future Work

Although the APPTI framework and suggested analysis show great promise, there are still a number of interesting research possibilities that could be pursued. First, large-scale and multilingual datasets, including generative and instruction-tuned LLMs like GPT-style and Llama-based models, should be evaluated in future research in addition to SST-2. These models show distinct memorizing tendencies and could uncover novel relationships between prompt-based attacks, robustness, and privacy leaking. This research conducted a comprehensive comparative investigation of privacy-preserving transformer inference under realistic adversarial threat models. By combining information from twenty one foundational studies covering differential privacy,

cryptographic secure inference, certified robustness, membership inference, backdoor attacks and adversarial prompt manipulation, the work created a unified analytical framework for assessing model vulnerabilities in cloud-based MLaaS environments. Five sample architectures which are DistilBERT, BERT-base, RoBERTa-base, RoBERTa-Large and a DeBERTa-v3 + RoBERTa ensemble were improved and evaluated against MIA and clean-label backdoor attacks using the SST-2 dataset. The results demonstrate that inference-time defenses significantly improve privacy and robustness without compromising the scalability or utility of the model. These adaptive controllers would enable LLM systems to respond to evolving threats with intelligence.

Finally, when employing APPTI-like systems on an industrial scale, there are issues with cross-platform compatibility, technical viability and regulatory reporting. Future studies should examine how inference-time privacy and robustness defenses interact with cloud administration, A/B testing, real-time logging and legal frameworks such as GDPR, HIPAA and the EU AI Act. These problems need to be fixed in order to transform research-grade protections into solutions that are prepared for manufacturing.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.
2. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191.
3. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & McDaniel, P. (2021). Extracting training data from large language models. USENIX Security Symposium, 2633–2650.
4. Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2022). Cryptographic protocols for secure machine learning. Journal of Cryptology, 35(2), 1–32.
5. Chen, J., Zhang, R., Wang, S., & Ren, K. (2023). TEE-based confidential inference for transformers. Proceedings of the IEEE Symposium on Security and Privacy.
6. Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. International Conference on Machine Learning, 1310–1320.
7. Li, B., Xie, C., & Li, Y. (2023). SoK: Certified robustness for deep neural networks. Proceedings of the IEEE Symposium on Security and Privacy.
8. Lu, S., Zheng, Z., Zhang, K., & Chen, X. (2023). Bumblebee: Secure two-party transformer inference. Proceedings of the USENIX Security Symposium.
9. Lyu, L., He, X., & Li, Y. (2020). Differentially private representation learning. Advances in Neural Information Processing Systems (NeurIPS).
10. Mireshghallah, F., Tople, S., & Shokri, R. (2021). Privacy regularization: Protecting user data during inference. Proceedings of the AAAI Conference on Artificial Intelligence.
11. Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. IEEE Symposium on Security and Privacy, 739–753.
12. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. International Conference on Learning Representations (ICLR).

13. Salman, H., Yang, G., Zhang, H., Zhang, C., Hsieh, C.-J., & Zhang, P. (2020). Provably robust deep learning via adversarially trained smoothed classifiers. Advances in Neural Information Processing Systems (NeurIPS).
14. Shan, S., Zhang, Y., Chen, H., & Li, B. (2024). Nightshade: Clean-label prompt poisoning for multimodal models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
15. Shen, Y., Yu, T., & Jin, H. (2024). Jailbreaking large language models via adversarial prompts. Proceedings of the ACM Conference on Computer and Communications Security (CCS).
16. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. IEEE Symposium on Security and Privacy, 3–18.
17. Sun, L., Xu, R., & Liu, J. (2024). Prompt filtering for safe large language model deployment. Proceedings of the AAAI Conference on Artificial Intelligence.
18. Wei, J., Zhou, X., & Zhang, Y. (2024). Detecting and mitigating jailbreak attacks in language models. Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).
19. Yu, T., Li, M., & Huang, L. (2024). Adversarial attacks on safety filters of LLMs. Proceedings of the ACM Workshop on Artificial Intelligence and Security.
20. Zeng, J., Lin, H., & Chen, Y. (2023). Narcissus: Clean-label backdoor attacks in text classification. Proceedings of the USENIX Security Symposium.
21. Zhang, H., Wang, X., & Ren, K. (2024). Non-interactive secure inference for transformers. Proceedings of the IEEE Symposium on Security and Privacy.