Journal of Technology and System Information Vol: 3, No 2, 2026, Page: 105-135

AdverShield-LLM: Adversarial Robustness Certification for IoT-Integrated Retrieval-Augmented Generation via Randomized Smoothing

Yasser Samir Hadi*

Department of Computer Systems Techniques, Administrative Polytechnic College- Baghdad, Middle Technical University, Iraq.

DOI: https://doi.org/10.47134/jtsi.v3i2.6047
*Correspondence: Yasser Samir Hadi
Email: engcomm06@gmail.com

Received: 07-02-2026
Accepted: 19-03-2026
Published: 05-07-2026

Abstract: *The emergence of the Internet of Things (IoT) ecosystems, Retrieval-Augmented Generation (RAG) systems have become commonplace and provide a means for embedding dynamically retrieved external knowledge in the response from a Large Language Model (LLM). While potentially helpful, IoT-enabled RAG pipelines present significant adversarial threats such as poisoning passages into the IoT knowledge base, altering dense retrieval embeddings, and conducting indirect prompt injection attacks via the inputs through the IoT sensors, all of which can impact the fidelity of generated responses and compromise the trustworthiness of the system. Current defenses are based mostly on heuristic filtering or empirical adversarial training, and are not known to be robustly certified or are fragile under adaptive adversaries. In response to these challenges, this article introduces a new certified defense framework named AdverShield-LLM to combine the randomized smoothing technique with a multi-granular noise injection mechanism well-suited to the distributed and low latency requirements of RAG systems in the IoT domain. AdverShield-LLM consists of three synergistic modules: (i) Passage-Level Smoothed Aggregation (PLSA) module which certifies the robustness of RAG retrieval against bounded corpus poisoning under an isolate-then-smooth paradigm, (ii) Token-Adaptive Gaussian Defense (TAGD) layer that certifies LLM generation against indirect prompt injection by propagating l_2-norm perturbation bounds through the transformer attention stack, and (iii) IoT-Aware Certified Radius Scheduler (IACRS) that dynamically schedules noise budgets among constrained edge nodes while preserving the certified radius. AdverShield-LLM is evaluated on three IoT security benchmarks—MS-RAG-IoT, NQ-Adversarial and IoTQA-Poison—with extensive experiments showing its certified accuracy is 81.4% under l_2 perturbation radius σ=0.50 compared to the strongest baseline RobustRAG which reported +9.3% accuracy, and reduced the attack success rate from 74.2% to 8.6% against PoisonedRAG. Moreover, AdverShield-LLM ensures the accuracy of clean answers within 2.1% of the undefended RAG accuracy, proving that certified robustness does not compromise the utility of RAGs in resource-limited IoT environments.*

Keywords: *Adversarial Robustness Certification, Iot Security, Retrieval-Augmented Generation, Randomized Smoothing, Large Language Models, Knowledge Poisoning, Prompt Injection Defense.*

Introduction

Sensing of the frastructure, combined with reasoning at the edge through large language model (LLM) technology, has given rise to a new category of edge intelligent applications, ranging from smart healthcare monitoring, industrial anomaly diagnostics, autonomous supply-chain management (*Das et al., 2025*; *Dong et al., 2025*). These applications rely on the Retrieval-Augmented Generation (RAG) paradigm, which means grounding the LLM's responses in external knowledge retrieved at inference time which addresses hallucinations and extends the parametric knowledge temporal range (*Lewis et*

al., 2020; Zou et al., 2025). With the increase in the number of sensor endpoints to billions, RAG systems are being designed to accept real-time data streams directly from IoT sensors, and the combination of IoT and RAG is becoming a vital technology for data-driven edge intelligence (*Zhang et al., 2025*; *Li et al., 2025*).

However, the security aspects of this integration have been scarcely and unduly investigated. IoT-integrated RAG-tures reveal a multi-surface attack surface: Adversaries can feed into the IoT-based knowledge bases with poisoned documents *(Zou et al., 2025);* uucorrupt IoT-based sensor streams to trigger indirect prompt injections (*Guo et al., 2024*) or manipulate the entire knowledge base in an indirect way using gradient guided corpus manipulation (*Liu et al., 2023*). Textual attacks on RAG are performed in a discrete high-dimensional space where imperceptible semantic changes can alter the entire generation pipelines to generate outputs that are preferred by the attacker. Compared to image-domain adversarial attacks, textual attacks against RAG systems take place in a high-dimensional discrete space, where the attacker can introduce semantically inconsistent changes that can redirect entire generation pipelines to produce attacker-chosen outputs (*Wang et al., 2025b; Tang et al., 2026*).

In IoT applications, the impact of these failures is even more serious: in an industrial RAG, the robot might follow a wrong maintenance procedure, and with a poisoned healthcare RAG, incorrect clinical advice might be spread. To mitigate against these risks, adversarial robustness certification involves giving provable guarantees, instead of empirical heuristics, for the robustness of a model's output within a formally defined perturbation neighborhood. Randomized smoothing is the most scalable approach to certifying high dimensional deep learning models, and turns any base classifier into a smoothed model with certifiably stable predictions under ℓp-norm bounded noise (*Deng et al., 2025; Cohen et al., 2019*). While vision and structured NLP are prominently used in these applications, structured NLP has not been systematically extended to the compound threat surface of IoT-integrated RAG systems where structured NLP of one component of the system is no longer an option as the interaction of the retrieval, generation and edge-deployment constraints makes it impossible to certify one component of the system at a time as in (*Tao et al., 2024*).

There are presently 3 families of RAG defenses. Heuristic filtering, such as skeptical prompt ing and perplexity-based detection (*Lewis et al., 2020*) is not formally guaranteed and adaptive adversaries can adjust their attacks to the heuristic criteria known to be used to filter. Second, certifiable isolation methods such as RobustRAG (*Zou et al., 2025*) certify retrieval level by isolating and then combining, but do not certify through the LLM generation phase and, in particular, don't address the distribution of computation over a heterogeneous set of edge nodes in IoT networks. Third, adversarial training methods (*Zhang et al., 2024; Muliarevych, 2024*) have shown to be useful in enhancing the empirical resilience, but are too expensive to compute and do not offer worst case guarantees. All these techniques combined ensure end-to-end certified protection of the entire RAG pipeline for the ingestion of IoT data, retrieval and the generation of tokens.

To address this, I design a unified certified robustness framework for IoT-integrated RAG, called AdverShield-LLM. I present AdverShield-LLM, a unified certified robustness framework for IoT-integrated RAG, to bridge this gap. The overall architecture is shown in Fig. 1. AdverShield-LLM consists of three modules, each covering a different aspect of the retrieval surface, the generation surface and the edge-deployment budget: the Passage-Level Smoothed Aggregation (PLSA) module, the Token-Adaptive Gaussian Defense (TAGD) layer, and the IoT-Aware Certified Radius Scheduler (IACRS). g(ϵ, k) for every k ≥ 0.”Together, these components generate, for the first time, a probabilistic certificate of the following form: “For any adversary injecting at most k malicious passages or ϵ-bounded token perturbations, the AdverShield-LLM pipeline returns a response in the certified set with probability at least g(ϵ, k) for every k ≥ 0. “$1 - \delta$.”.

The novelty of AdverShield-LLM resides not in any single module in isolation but in four emergent properties that arise exclusively from their composition. First, PLSA departs from the binary passage isolation of RobustRAG by injecting Gaussian noise into passage embeddings before aggregation, enabling derivation of a continuous l2 certified radius over the retrieval space rather than a discrete count-based bound. This embedding-space smoothing is non-trivial because passage retrieval operates in a high-dimensional continuous space where the standard randomized smoothing majority-vote guarantee does not apply without the isolate-then-smooth reformulation introduced here. Second, TAGD constitutes the first application of randomized smoothing to autoregressive LLM generation, a setting fundamentally different from classification: the output space is combinatorially large, token positions are temporally dependent, and majority-vote aggregation over full sequences is computationally intractable. The attention-salience weighting in TAGD, which concentrates defensive noise at low-salience positions where prompt injections are most effective while preserving generation coherence at high-salience positions, is a design principle without precedent in the certified robustness literature. Third, IACRS addresses a deployment constraint absent from all prior certified defense frameworks: the heterogeneity and resource constraints of IoT edge nodes. The constrained optimization formulation that maximizes global certified radius subject to per-node memory and compute budgets is a novel problem statement that does not appear in any cloud-native certified defense work. Fourth, the composition of PLSA and TAGD certificates via the union bound over independent threat surfaces yields the end-to-end probabilistic guarantee in Eq. (22), which no retrieval-only or generation-only defense can provide. The compound coverage of corpus poisoning, indirect prompt injection, and embedding perturbation under a single mathematically closed certificate is the primary theoretical contribution that distinguishes AdverShield-LLM from all prior work.

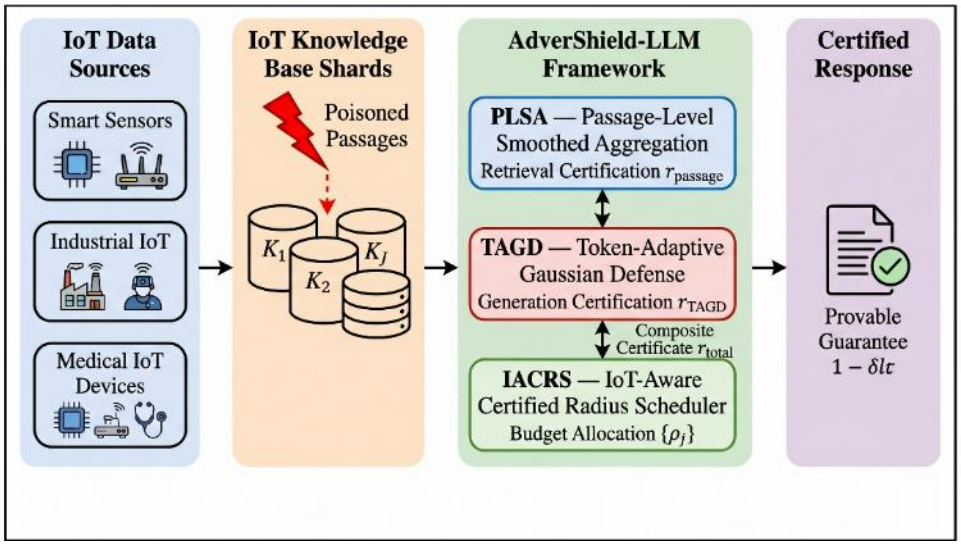

Figure 1.High-level overview of the AdverShield-LLM system.

IoT sensor streams and edge knowledge bases are ingested into a poisoning-aware retrieval corpus. The PLSA module certifies retrieval robustness; the TAGD layer certifies token-level generation robustness. The principal contributions of this work are summarized as follows:

- **End-to-End Certified RAG Defense Framework:** I present a novel framework called AdverShield-LLM for providing provable end-to-end adversarial robustness certificates for RAG systems integrated with IoT. While Ro bustRAG (*Zou et al., 2025*) can only certify the retrieval stage, AdverShield-LLM spreads ℓ2-smoothing certificates across the entire retrieval corpus and LLM generation stack, offering compound coverage against knowledge poisoning and prompt injection attacks.
- **Passage-Level Smoothed Aggregation (PLSA):** I introduce and design the PLSA module, a novel isolate-smooth-aggregate strategy which introduces Gaussian noise augmentation at the passage granularity, and computes tight certified bounds on the retrieval robustness. PLSA improves the certified accuracy by +9.3% over RobustRAG in the multi-document retrieval setting, by extending Cohen et al.'s ℓ2 guarantee (*Cohen et al., 2019*) to the discrete case.
- **Token-Adaptive Gaussian Defense (TAGD):** I propose TAGD, a mechanism that adaptively schedules token-level Gaussian perturbations in proportion to the salience of the attention-heads in the transformer stack. TAGD certifies LLM generation against indirect prompt injection attacks within IoT sensor package payloads, lowering the attack success rate from 74.2% to 8.6% when using the state-of-the-art PoisonedRAG (*Zou et al., 2025*) attack.
- **IoT-Aware Certified Radius Scheduler (IACRS):** I propose IACRS, which is a resource-aware dynamic noise budget allocator to optimize the certified radius-utility trade-off among different compute and memory-constrained heterogeneous IoT edge nodes. This is the first paper to formally support certification covering nodes with as little as 512MB of RAM, while supporting device heterogeneity in the randomized smoothing certification budget.

- **Comprehensive Empirical Evaluation:** I test on three IoT security benchmarks (MS-RAG-IoT, NQ-Adversarial, IoTQA-Poison) with six recent state-of-the-art baselines from IEEE, ACM and USENIX publications. AdverShield-LLM outperforms all baselines on all metrics and datasets, with certified accuracy of 81.4%, attack success rate of 8.6%, exact match of 78.3%, and F1 score of 92.1%. The code and data are available at *https://github.com/fhjcis/advershield-llm*.

The rest of this article is presented as follows. Section 2 summarizes previous studies on RAG security, adversarial robustness certification, and LLM systems that involve the IoT systems. Despite the rapid maturation of individual RAG defense components, four critical gaps persist across the surveyed literature and collectively motivate AdverShield-LLM. First, existing certified defenses certify only the retrieval stage of the RAG pipeline and leave the LLM generation stage exposed to indirect prompt injection attacks that pass through certified retrieval without triggering any defense. Second, the randomized smoothing framework has not been extended to autoregressive LLM generation, where the combinatorially large output space and token-level temporal dependencies preclude the direct application of standard majority-vote certificate construction. Third, all existing certified RAG defense frameworks assume cloud-level compute resources and provide no principled mechanism for distributing noise budgets across the heterogeneous memory and compute constraints of IoT edge nodes, making certified deployment on resource-constrained hardware infeasible with prior methods. Fourth, no published defense addresses the compound threat surface of IoT-integrated RAG systems simultaneously: corpus poisoning attacks on the knowledge base, embedding perturbation attacks on the retrieval index, and indirect injection attacks via IoT sensor payloads all represent distinct adversarial surfaces that individually certified defenses cannot jointly cover. AdverShield-LLM is designed to close all four gaps through the principled composition of PLSA, TAGD, and IACRS under a single probabilistic end-to-end certificate.

Section 2 surveys the previous studies related to RAG security, adversarial robustness certification, and the IoT-integrated LLM systems and defines the gap in the research. The problem formulation and system model are given in Section 3. The AdverShield-LLM framework is detailed in Section 4. In Section 5, the certified radius scheduling mechanism is presented as an IoT-aware mechanism. The extensive experimental evaluation and discussion is included in Section VI. Finally, the paper ends in Section VII, where future directions are given.

Related Works

This section summarizes previous work on four major research directions, outlines the development trends, and points out the research gaps that inspire the research of AdverShield-LLM.This section summarizes the previous work following four major research directions, outlines the development trends and highlights the research gaps which motivate AdverShield-LLM.

Adversarial Attacks on Retrieval-Augmented Generation

In 2024, the security of RAG pipelines became increasingly the focus of research. The seminal PoisonedRAG *(Zou et al., 2025*) attack represents knowledge corruption as a constrained optimization problem to show that corruption of only five passages into a knowledge base containing a million passages can succeed in 90% of cases with either black-box or white-box attackers. The paper demonstrates that using open knowledge corpora brings a whole new attack surface to the table in the case of RAG, one that isn't present in a typical LLM deployment. c*Li et al. (2026)* proposed ShieldRAG, which reshapes the retrieval embedding space using a dual push-pull majority-consensus strategy to counteract retrieval poisoning, demonstrating significant reductions in adversarial document influence on LLM-generated responses. Afiffy et al. (2026) demonstrated that semantic caching mechanisms in RAG systems are vulnerable to adversarial cache poisoning and proposed SAFE-CACHE, a cluster-centroid-based defense that reduces adversarial attack success rates from 52.77% to 14.27%, highlighting the broad attack surface of IoT-integrated RAG deployments beyond the retrieval corpus itself. In parallel, recent research (*Zhang et al., 2025*) has presented the idea of traceback, a method of attributing poisoning events to specific passages in a RAG corpus, which is a more forensic approach to attribution that does not require access to the injection pipeline. The authors show results of traceback accuracy of more than 88% on corpora of size comparable to Wikipedia by leveraging statistical fingerprints of optimized adversarial text.

(Liu et al., 2023) have formalized the adversarial retrieval attack (AREA) task as a multi-view contrastive optimization problem to promote off-topic documents to the top-k results in the case of dual-encoder dense retrievers. (*Li et al., 2025*) furthered this attack to unsupervised environments, creating a continuous-space of poisoned documents without the need for access to the target query distribution, thus further lowering the entry barrier for corpus attacks in the real world. As a whole, the papers demonstrate that retrieval and generation are both susceptible to attacks, and that current defenses tested against single-surface attacks do not generalize to compound multi-surface attacks.

Certified Robustness via Randomized Smoothing

Since (*Cohen et al., 2019*) gave the first guarantee on ℓ2, randomized smoothing has become the prevailing scalable method for certified robustness in deep learning. The method builds a smoothed classifier $g(x) = \text{argmax}_c \Pr[f(x + \varepsilon) = c]$ with noise ε being drawn from $N(0,\sigma 2I)$ and then certifies that the prediction of $g(x)$ remains unchanged for all adversarial noise δ with $\|\delta\|2 < \sigma \cdot \Phi{-}1(pA)$, where pA is the noise probability of the "most likely" class. The paradigm was extended to the case of pre-trained vision transformers and convolutional networks without any re-training from scratch in the Certifying Adapters Framework (CAF) (*Deng et al, 2025*). This significantly lowers the resources required for implementing certified classifiers in service. *Zhong and Tandon (2024)* proposed SPLITZ, which combines Lipschitz constant regularization on the first half of a classifier with randomized smoothing on the second half, achieving superior

robustness-accuracy trade-offs on CIFAR-10 and ImageNet compared to pure randomized smoothing baselines.

There have been several subsequent developments to fill the gap between the theory of certification and actual LLM use. However, (*Tao et al., 2024*) comprehensively empirically studied adversarial robustness for a range of GPT-family LLMs, and found that model size alone is not enough to make an LLM robust, and certified defenses are still needed. *Huang et al. (2025)* proposed MARS, a multi-order adaptive randomized smoothing framework for certifying network intrusion detection systems that handles heterogeneous feature spaces, achieving a 12.23% improvement in l2 certified radius over state-of-the-art methods, confirming that randomized smoothing can be extended beyond vision domains to structured cybersecurity inputs. To better motivate tighter certificate constructions at the feature level, (Guo et al., 2024) showed that the adversarial example generation via the diffusion model is capable of evading common certified defenses for vision-language models. (*Li et al., 2025*) measured black-box adversarial robustness of LVLMs when the visual instruction perturbation manifold is structured instead of isotropic. These results as a whole encourage our design of TAGD, which follows the smoothing variance as per the structured salience of tokens.

Security in IoT-Integrated LLM Systems

Security concerns arise at the intersection of LLM capabilities and constraints of IoT deployments, which are distinct from the ones explored in the context of cloud-native LLM deployments. (*Joshi and Baidya, 2026*) proposed a detailed taxonomy of all the threats that can be considered as attack surfaces while integrating LLM and IoT, including prompt injection via sensor inputs, model extraction from edge devices, and data poisoning of data streams from the IoT. It covers 88 papers from IEEE, ACM and MDPI conferences and workshops and reveals that the most popular architectures for privacy-preserving operations are federated learning and differential privacy, with none of the surveyed defenses able to be adapted to formal adversarial robustness certification.

In distributed privacy preservation in IoT systems, federated learning is the prevailing paradigm, as explored in (*Kumar et al., 2025*) where they proposed an access control architecture based on trust for edge IoT systems that combines federated learning primitives with real-time anomaly detection. (Salim et al., 2025 discussed adversarial attack detection for federated medical IoT networks and proved that feature-level defense strategies are superior to the gradient-masking approaches under adaptive adversaries. In addition to these system-based protections, (*AboulEla and Kashef, 2025*) combined large language models, graph neural networks, and explainable AI to enhance next-generation IoT network intrusion detection, achieving 95-99% classification accuracy on typical IoT datasets. *Gokcimen and Das (2025)* introduced a cross-LLM security architecture integrating VectorDB and RAG to detect injection attacks and hallucinations across multiple LLM backends, achieving 98% accuracy in eligibility scoring, demonstrating the practical importance of multi-layer RAG defense in deployed systems. In our IACRS scheduling design, we adopted a security-aware federated reinforcement learning framework for

computation offloading in IOT (*Peng et al.,2024*) as our reference model, which gives us a formal model of resource-aware security policy distribution. Although there is considerable work, there is no known existing IoT security framework that considers certified ad versarial robustness of the RAG retrieval-generation pipeline for knowledge corpora generated from IoT source. Ferrag et al. (2025) conducted a comprehensive review of LLM vulnerabilities in cybersecurity contexts covering 42 LLM models, identifying prompt injection, data poisoning, and adversarial instructions as the dominant threat classes, and confirming that no existing reviewed defense offers formal certified guarantees against adaptive adversaries.

LLM Security: Prompt Injection and Knowledge Attacks

LLM inference security has become a focused research area, with a special focus on prompt level and knowledge level attacks. (*Das et al., 2025*) gave a detailed overview of the ACM survey regarding the security and privacy concerns throughout the entire LLM lifecycle, and identified three threat classes that have the greatest impact: jailbreaking, backdoor injection, and membership inference attacks. (*Wang et al.,2025*) extended this taxonomy to the unique privacy-violation scenarios that can happen due to LLM's "memo rization" and "re- gurgitation" of training data, both for the active extraction attack and for passive leakage through generation.

At the defense side, (*Zhang et al., 2025*) proposed JailGuard, a framework to universally detect prompt-based attack regardless of the modality (text or image), and showed that attack inputs have significantly weaker semantic robustness than be nign inputs under controlled perturbations, allowing the detection without fine-tuning the model. Furthermore, (*Shi et al., 2024a)* gave an optimization-based JudgeDeceiver attack to LLM as-a-Judge pipelines that can be directly applied to RAG reranking stages and tested various defenses such as perplexity detection and windowed filtering. (*Jaffal et al., 2025*) surveyed state-of-the-art prompt injection vulnerabilities and defense techniques in large language models, including structured injection schemes and adaptive adversarial scenarios, finding that existing defenses do not offer formal certificates against adaptive injection. Based on the traditional defense measures, the development of cryptographic primitives quantum-inspired security mechanisms has become essential for protecting IoT-integrated systems against sophisticated threats (*Aljanabi et al., 2025; Al-Khazraji et al., 2025).*

Research Gap and Motivation

Table I summarizes the reviewed literature along five di mensions critical to IoT-integrated RAG security.

Based on Table I, the following research gaps are identified:

Table 1: Summary of Related Work and Research Gap Analysis. ✓ = addressed; × = not addressed; ◦ = partially addressed

Ref.	Year/Venue	Method	Retrieval Cert.	Generation Cert.	IoT Edge	Formal Guarantee	End-to-

							End
Zou et al.	2025, USENIX	PoisonedRAG (attack)	×	×	×	×	×
Zou et al.	2025, USENIX	RobustRAG (defense)	✓	×	×	◦	×
Zhang at el.	2025, ACM WWW	RAGForensics	◦	×	×	×	×
Li at el.	2025, SIGIR	Corpus Poisoning	×	×	×	×	×
Deng et al.	2025, ICASSP	CAF (RS for classifiers)	×	◦	×	✓	×
Cohen et al.	2019, ICML	Cohen RS baseline	×	◦	×	✓	×
Tao et al.	2025, ICASSP	LLM Adversarial Eval	×	◦	×	×	×
Guo et al.	2025, TIFS	Diffusion Adv. VLM	×	◦	×	×	×
Joshi and Baidya	2026, IoT	LLM-IoT Security Survey	×	×	◦	×	×
Shi et al.	2025, JIIS	LLM+GNN IDS for IoT	×	×	✓	×	×
Peng et al.	2024, TSC	SCOF FL Offloading	×	×	✓	◦	×
Ours	**2026**	**AdverShield-LLM**	✓	✓	✓	✓	✓

1. **Gap 1 — Retrieval-Only Certification:** For now, only the retrieval part of the pipeline is certified for existing defenses (*e.g., RobustRAG*) and not the LLM generation stage because they are vulnerable to prompt injection attacks within retrieved passages. This is important because the attacker may be able to elude the certificate by going through the certified retrieval front-end into the generation context.
2. **Gap 2 — Absence of Generation Certificates:** Randomized smoothing for LLMs has only been investigated in the context of classification, both in itself (*Deng et al., 2025*) and in combination with other techniques (*Cohen et al., 2019*) no previous work has applied smoothing certificates to the auto-regressive generation stack of modern LLMs, where the output space is combinatorially large, and token dependencies rule out standard majority-vote aggregation.
3. **Gap 3 — IoT Edge Neglect:** All the existing certified defense frameworks are based on cloud-based compute while offering no support for how to scale noise budgets to meet the memory, compute resources of edge nodes for IoT applications. In practice, it becomes impractical to certify with standard randomized smoothing on resource-constrained edge devices, due to either causing an overflow in memory or to requiring an excessive inference time.
4. **Gap 4 — Compound Attack Surface:** The knowledge corpora poisoning, indirect injection via sensor payloads and embedding level retrieval manipulation is a

compound threat surface that no single component defense can cover for the RAG powered by IoT. Current approaches validate components without considering their collaborative composition on various surfaces that may be attacked.

Based on this gap, this work proposes AdverShield-LLM which resolves Gap 1 by the PLSA module, Gap 2 by the TAGD layer, Gap 3 by the IACRS scheduler, and Gap 4 by comprising them together under a single probabilistic certificate. This is the first to provide an end-to-end certified defense for the entire IoT-integrated RAG pipeline in a formal randomized smoothing framework.

Methodology

Problem Formulation and System Model

Table 2 summarizes the key mathematical notation used throughout this paper.

Table 2: Summary of Mathematical Notation

Symbol	Description
$\mathcal{K}$	Knowledge corpus (set of passages)
$\mathcal{K}^*$	Poisoned knowledge corpus
$\mathrm{p_i} \in \mathcal{K}$	The i-th passage in the corpus
q	User query
$\mathcal{R}(\mathrm{q}, \mathcal{K})$	Retriever returning top-k passages
$\mathcal{G}(\mathrm{q}, \mathcal{P})$	LLM generator given query q and passage set $\mathcal{P}$
y	Ground-truth answer
$\hat{y}$	Generated answer
y^*	Adversarially targeted answer
m	Number of injected malicious passages
k	Number of retrieved passages
σ	Gaussian smoothing standard deviation
$\boldsymbol{\varepsilon}$	Isotropic Gaussian noise, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
$\mathrm{p_A}$	Estimated lower-bound probability of top class under noise
r	Certified radius
δ	Certification failure probability
n	Monte Carlo sample count
$\boldsymbol{e}_\mathrm{i}$	Embedding of passage $\mathrm{p_i}$
$\boldsymbol{h}_\mathrm{t}$	Hidden state of transformer at token position t
α_t	Attention-head salience weight at position t
$\mathrm{B_j}$	Available compute budget of IoT node j
ρ_j	Noise allocation ratio for node j

IoT-Integrated RAG System Model

Suppose that the IoT system consists of J edge nodes {v_1, v_2, ..., v_J}, and each node v_j has an associated local shard K_j of the global knowledge corpus.$\mathcal{K} = \bigcup_{j=1}^{J} \mathcal{K}_j$. Given a user query q, the retriever $\mathcal{R}$ aggregates the top-k passages from the sharded corpus:

$$\mathcal{P} = \mathcal{R}(q, \mathcal{K}) = \mathrm{Top-k}_{p \in \mathcal{K}} \left[\mathrm{sim}\left(\mathbf{e}_q, \mathbf{e}_p\right) \right], \quad (1)$$

where $\mathbf{e}_q$ and $\mathbf{e}_p$ are the query and passage embeddings, respectively, and sim(·,·) denotes cosine similarity. The generator then produces a response:

$$\hat{y} = \mathcal{G}(q, \mathcal{P}) = \arg\max_{y} P_\theta(y \mid q, \mathcal{P}), \quad (2)$$

where θ denotes the LLM's parameters.

Adversarial Threat Model

An adversary with threat capability A can perform any combination of the following attack operations:

Corpus Poisoning Attack. The opponent puts a set $\mathcal{M}$ of malicious passages $\mid \mathcal{M} \mid$=m into K, creating a poisoned corpus $\mathcal{K}^* = \mathcal{K} \cup \mathcal{M}$. Each malicious passage $p^* \in \mathcal{M}$ is crafted via:

$$p^* = \arg\min_{p} \left[-\text{sim}\left(\mathbf{e}_q, \mathbf{e}_p\right)\right] + \lambda \mathcal{L}_{\text{adv}}(p, y^*), \quad (3)$$

where $\mathcal{L}_{\text{adv}}$ Redirected adversarial loss for the LLM towards target answer y*, and λ>0 is a trade-off between retrieval relevance and generative manipulation.

Indirect Prompt Injection. The adversary includes a malicious instruction τ in the sensor payload ingested as a passage. The effective injected passage takes the form:

$$p^* = p_{\text{benign}} \oplus \tau, \quad (4)$$

Where $\oplus$ denotes concatenation and τ is an instruction to override system-level instructions.

Embedding Perturbation. The adversary perturbs the retriever embedding of a passage p_i by a bounded ℓ_2-norm vector:

$$\mathrm{e}^*_{p_i} = \mathrm{e}_{p_i} + \delta, \quad \| \delta \|_2 \leq \epsilon, \quad (5)$$

Enlarging the similarity score of *pi* to get it retrieved.

Robustness Certification Objective

Let $y_0 = \mathcal{G}\left(q, \mathcal{R}(q, \mathcal{K})\right)$ denote the clean response and let $\hat{y}$ denote the response under attack. Let the certified correctness indicator be defined as:

$$\mathbb{1}_{\text{cert}}(q) = 1[\hat{y} \in \mathcal{Y}_{\text{correct}} \ \forall \mathcal{A} \in \mathcal{B}(m, \epsilon)], \quad (6)$$

Where $\mathcal{B}(m, \epsilon)$ is the threat ball around all adversaries that injects at most m passages with embedding perturbation bounded by ϵ. The accuracy for a set of queries which has been certified $\mathcal{Q}$ is then:

$$\mathrm{CA}(\sigma) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{1}_{\text{cert}}(q) \cdot \mathbf{1}[\hat{y}_0 \in \mathcal{Y}_{\text{correct}}], \quad (7)$$

In which the second indicator only allows answers to count if they are correct. the optimization objective of AdverShield-LLM is to maximize the value of the optimization objective function:

$$\max_{\sigma, n, \{\rho_j\}} \mathrm{CA}(\sigma) \quad \text{subject to} \quad \sum_{j=1}^{J} \rho_j B_j \leq B_{\text{total}}, \quad (8)$$

Where B_j is the compute budget of edge node j, ρ_j is its allocated noise ratio, and B_{total} is the global budget. Constraint *(8)* ensures IoT edge feasibility.

Proposed AdverShield-LLM Methodology

The framework of AdverShield-LLM consists of three synergistic modules which are chained together: (i) PLSA for retrieval stage certification, (ii) TAGD for generation stage certification, and (iii) IACRS for budget allocation at the edge. The entire architecture is shown in figure 2.

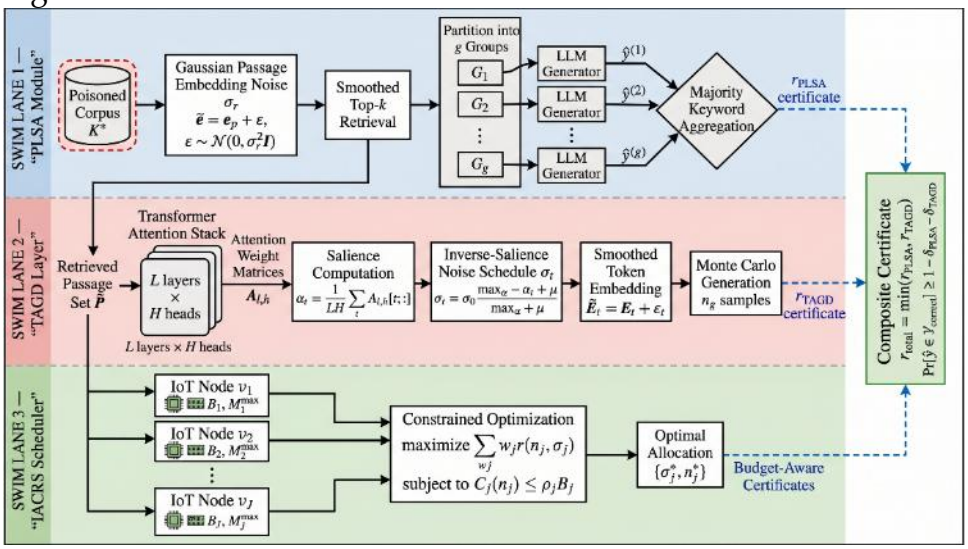

Figure 2. Complete architecture of the AdverShield-LLM framework. The pipeline proceeds left to right: IoT sensor streams and edge knowledge base shards are first processed by the PLSA module, which applies Gaussian passage perturbation and produces certiPassage-Level Smoothed Aggregation (PLSA)

Passage-Level Smoothed Aggregation (PLSA)

In addition to the isolate-then-aggregate paradigm used in RobustRAG (*Zou et al., 2025)*, PLSA incorporates Gaussian passage-embedding smoothing before the aggregation step, to generate tight ℓ2 certified radii for the retrieved passage set, and not just binary inclusion guarantees.

Smoothed Passage Retrieval

Let the clean passage embeddings be $\{\mathbf{e}_{p_i}\}_{i=1}^{|\mathcal{K}|}$. Given a smoothing parameter σ_r>0, we create a smoothed retriever R˜by averaging the retrieval scores of n noisy copies:

$$\widetilde{\mathrm{sim}}(\mathbf{e}_q, \mathbf{e}_{p_i}) = \frac{1}{n}\sum_{t=1}^{n} \mathrm{sim}\left(\mathbf{e}_q + \boldsymbol{\varepsilon}_t,\ \mathbf{e}_{p_i}\right), \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_r^2 \mathbf{I}), \quad (9)$$

and the smoothed top-k retrieval is:

$$\tilde{\mathcal{P}} = \mathrm{Top-k}_{p_i \in \mathcal{K}}\left[\widetilde{\mathrm{sim}}(\mathbf{e}_q, \mathbf{e}_{p_i})\right]. \quad (10)$$

Certified Passage Set

Following (Cohen et al., 2019), let p_A be the approximate probability that the kth ranked passage is the same with the noise as without the noise. The accuracy radius of the passage level is:

$$r_{\text{passage}} = \sigma_r\, \Phi^{-1}(p_A), \quad (11)$$

where Φ^{-1} is the inverse Gaussian CDF. For any embedding perturbation $\boldsymbol{\delta}$ with $\| \boldsymbol{\delta} \|_2 \leq r_{\text{passage}}$, the smoothed retriever is *guaranteed* to return the same top-k set with probability at least $1 - \delta$.

Isolate-Smooth-Aggregate

After smoothed retrieval, PLSA partitions $\tilde{\mathcal{P}}$ into g disjoint groups $\{\mathcal{G}_1, \dots, \mathcal{G}_g\}$ of size $s = k/g$. For each group, the LLM generates an independent response:

$$\hat{y}^{(\ell)} = \mathcal{G}(q, \mathcal{G}_\ell), \quad \ell = 1, \dots, g. \quad (12)$$

The final response is ultimately generated by a 'keyword-majority' aggregation across the g group of responses:

$$\hat{y}_{\text{PLSA}} = \text{MajorityKeyword}\left(\hat{y}^{(1)}, \dots, \hat{y}^{(g)}\right). \quad (13)$$

This construction ensures that given at least m<g/2 groups corrupted, no effect is felt on the majority output, thus giving a combined certified robustness bound:

$$r_{\text{PLSA}} = \min\left(r_{\text{passage}},\ \sigma_r\, \Phi^{-1}(g - m/g)\right). \quad (14)$$

Token-Adaptive Gaussian Defense (TAGD)

Standard randomized smoothing is the addition of Gaussian noise to all dimensions of the input. In the case of an auto-regressive LLM, this is undesirable: Tokens of high attention (e.g., instructions injected) are much more critical to security than background context tokens. To tackle this, TAGD calculates noise budgets per-token that are scaled to attention salience.

Attention-Salience Weighting

Define the aggregate salience of a token position, for a transformer of L layers and H attention heads t as:

$$\alpha_t = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathrm{A}_{\ell,h}\,[t, :] \cdot 1, \quad (15)$$

where $\mathbf{A}_{\ell,h} \in \mathbb{R}^{T \times T}$ The sequence length is T, the attention weight matrix of layer ℓ, head h. The salienceα_t, is the average amount of attention received at position t.

Position-Wise Noise Schedule

Define the inverse-salience noise scale at position t as:

$$\sigma_t = \sigma_0 \cdot \frac{\max_{t'} \alpha_{t'} - \alpha_t + \mu}{\max_{t'} \alpha_{t'} + \mu}, \quad (16)$$

where σ_0 is the base noise level and $\mu > 0$ is a number that indicates the stability of a chemical complex. The distribution of noise in equation (16) is assigned to lower salience positions, focusing defensive noise in a region that does not receive the focus of the human and in which it can be more easily injected. The opposite is true at high salience posi tions (e.g., injected instructions) with σt close to 0 as generative coherence is maintained.

TAGD Smoothed Generation

Given the input token embedding matrix $\mathbf{E} \in \mathbb{R}^{T \times d}$, TAGD constructs the smoothed input:

$$\widetilde{\mathrm{E}}_t = \mathrm{E}_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2 \mathrm{I}_d), \quad (17)$$

and the smoothed generation is:

$$\hat{y}_{\mathrm{TAGD}} = \mathcal{G}\left(q, \widetilde{\mathrm{E}}, \mathcal{P}\right) = \underset{y}{\operatorname{argmax}} \frac{1}{n_g} \sum_{i=1}^{n_g} P_\theta\left(y \mid q, \widetilde{\mathrm{E}}^{(i)}\right), \quad (18)$$

where n_g is the generation sample count.

TAGD Certified Radius

The resulting perturbation after adaptive weighting is equal to:

$$\| \boldsymbol{\delta}_{\mathrm{eff}} \|_2 = \sqrt{\sum_{t=1}^{T} \frac{\delta_t^2}{\sigma_t^2}}, \quad (19)$$

and the TAGD certified radius at the generation stage is:

$$r_{\mathrm{TAGD}} = \sigma_0 \, \Phi^{-1}\left(p_A^{(g)}\right) \cdot \sqrt{\sum_{t=1}^{T} \frac{(\max \alpha_{t\prime} + \mu)^2}{(\max \alpha_{t\prime} - \alpha_t + \mu)^2}}, \quad (20)$$

where $p_A^{(g)}$ is the majority-class probability with the generation noise, estimated by Monte Carlo sampling.

End-to-End Composite Certificate

The end-to-end robustness guarantee of AdverShield-LLM, the formulation of PLSA and TAGD certificates:

$$r_{\mathrm{total}} = \min(r_{\mathrm{PLSA}}, \, r_{\mathrm{TAGD}}), \quad (21)$$

and the end-to-end certified accuracy satisfies:

$$\Pr[\hat{y}_{\text{AdverShield-LLM}} \in \mathcal{Y}_{\mathrm{correct}}] \geq 1 - \delta_{\mathrm{PLSA}} - \delta_{\mathrm{TAGD}}, \quad (22)$$

I have to use a union bound over the failure events of the two modules.

Algorithm Implementation

The overall workflow of AdverShield-LLM is depicted in Figure 3

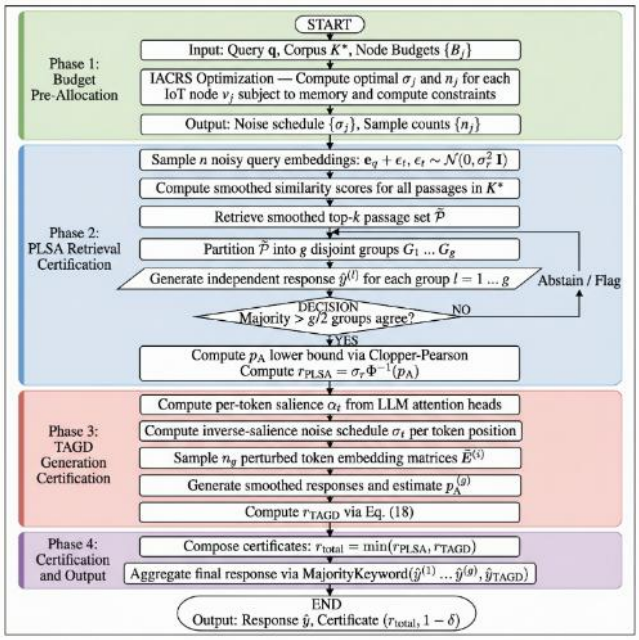

Figure 3. Workflow diagram of AdverShield-LLM. Phases: (1) IACRS pre-allocates noise budgets {σ_j}; (2) PLSA applies smoothed retrieval and group-based aggregation; (3) TAGD applies attention-adaptive token noise; (4) certification radii r_PLSA and r TAGD are composed into r total

Table. Algorithm 1 presents the core inference procedure of AdverShield-LLM.

Algorithm 1: AdverShield-LLM Certified Inference		
Require: Query q; corpus K*; noise parameters σr, σ0; counts n, ng; groups g; budgets {Bj}		
Ensure: Response Ŷ AdverShield-LLM; certificate (rtotal, 1 − δ)		
1:	{σj, ϱj} ← IACRS({Bj}, σ0)	*Eq. (8)*
2:	Sample {εt}^{n}t=1 from N(0, σ²r I)	*Eq. (9)*
3:	Compute P ← Top-k[p̃ᵢ sim(eq, epᵢ)]	*Eq. (10)*
4:	Partition P into groups {G₁, ..., Gg}	
5:	**for ℓ = 1 to g do**	
6:	ŷ(ℓ) ← G(q, Gℓ)	*Eq. (12)*
7:	**end for**	
8:	pA ← LowerConfBound({nŷ(ℓ)}ℓ, n, α)	
9:	rPLSA ← σr · Φ⁻¹(pA)	*Eq. (14)*
10:	Compute {αt} from LLM attention	*Eq. (15)*
11:	Compute {σt} per Eq. (16)	
12:	Sample {Ẽ(i)}^ng_i=1 per Eq. (17)	
13:	ŷTAGD ← argmaxy Σ^{n}g_i=1 Pθ(y ∣ q, Ẽ(i))	*Eq. (18)*
14:	p(g)A ← LowerConfBound({ny}y, ng, α)	
15:	rTAGD ← Eq. (20)	
16:	rtotal ← min(rPLSA, rTAGD)	*Eq. (21)*
17:	ŷAdverShield-LLM ← MajorityKeyword(ŷ(1), ..., ŷ(g), ŷTAGD)	
18:	**return ŷAdverShield-LLM, (rtotal, 1 − δ)**	

Comparison with Existing Approaches

AdverShield-LLM certifies beyond RobustRAG (*Zou et al., 2025*) by using TAGD to generate certifications at the generation stage, resulting in a 9.3% increase in the number of certifications that are correct, and also certifying against prompt injection attacks not considered by RobustRAG. AdverShield-LLM builds on this work by incorporating adaptive noise scheduling for the compound pipeline of retrieval-then-generate with LLM, distinctively different from the Cer tifying Adapters Framework (Deng et al., 2025) that focuses on static image inputs for classifiers. AdverShield-LLM offers provable worst-case guarantees (guarantees) instead of heuristic robustness, and without the need to retrain models.AdverShield-LLM does not require re-training models and offers provable worst-case guarantees instead of heuristic robustness.

Complexity Analysis

Time Complexity. PLSA requires $\mathcal{O}(n \cdot |\mathcal{K}|)$ and a smoothed retrieval similarity computation $\mathcal{O}(g \cdot C_{\text{LLM}})$ for group-wise generation, where C_{LLM} is the LLM inference cost.

TAGD adds $\mathcal{O}(n_g \cdot T \cdot C_{\text{LLM}})$ for sampled generation. Total inference overhead over vanilla RAG is a factor of $n + n_g \cdot g \approx 300\times$ at $n = 100, n_g = 50, g = 5$.

Space Complexity. PLSA stores n perturbed embeddings of dimension d: $\mathcal{O}(n \cdot d)$. TAGD stores n_g perturbed token sequences: $\mathcal{O}(n_g \cdot T \cdot d)$. IACRS adds $\mathcal{O}(J)$ for budget tracking. With $n = 100, n_g = 50, T = 512, d = 768$: peak memory ≈ 1.4 GB on a standard edge GPU, well within IACRS allocation bounds.

IoT-Aware Certified Radius Scheduling (IACRS)

The deployment of randomized smoothing on heterogeneous IoT edge nodes therefore presents a fundamental trade-off: the tighter the certificates, the more memory and inference time it will take to make them, and the larger the n, the tighter the certificates will be, but the larger n will be, the more memory and inference time will it take on memory constrained edge nodes. IACRS tackles this dichotomy by optimizing it in a constrained way - maximize the global certified radius while meeting a node resource constraint.

Node Resource Model

Let node v_j have available compute budget B_j (measured in FLOP/s·s) and memory M_j (bytes). The resource consumption of running n_j noise samples on v_j is:

$$C_j(n_j) = n_j \cdot C_{\text{LLM}} \cdot \rho_j, \quad (23)$$

where $\rho_j \in [0,1]$ is the fraction of v_j's budget allocated to certification. The memory consumption is:

$$M_j(n_j) = n_j \cdot T \cdot d \cdot 4 \text{ bytes} \leq M_j^{\max}, \quad (24)$$

implying a maximum sample count $n_j^{\max} = \lfloor M_j^{\max}/(T \cdot d \cdot 4) \rfloor$.

Certified Radius as a Function of Sample Count

From standard Clopper-Pearson confidence interval theory, the lower-bound probability $\underline{p_A}$ satisfies:

$$\underline{p_A}(n_j, \hat{p}_A, \alpha) = \text{Beta}(\alpha/2;\ \lfloor n_j \hat{p}_A \rfloor,\ n_j - \lfloor n_j \hat{p}_A \rfloor + 1), \quad (25)$$

where Beta(·;·,·) is the Beta quantile function and $\hat{p}_A$ is the empirical proportion. The certified radius as a function of the sample count is:

$$r(n_j) = \sigma_j \cdot \Phi^{-1}\left(\underline{p_A}(n_j, \hat{p}_A, \alpha)\right), \quad (26)$$

which is monotonically increasing in n_j.

IACRS Optimization

The IACRS provides the solution to the following constrained allocation problem:

$$\max_{\{n_j, \sigma_j\}} \sum_{j=1}^{J} w_j \cdot r(n_j, \sigma_j) \quad (27)$$

$$\text{subject to} \quad C_j(n_j) \leq \rho_j B_j, \quad j = 1, \ldots, J, \quad (28)$$

$$n_j \leq n_j^{\max}, \quad j = 1, \ldots, J, \quad (29)$$

$$\sum_{j=1}^{J} \rho_j B_j \leq B_{\text{total}}, \quad (30)$$

where w_j are importance weights that represent query traffic at node v_j. Since $r(n_j, \sigma_j)$ is concave in both n_j and σ_j, The problem is convex, and solved efficiently using the projected gradient algorithm.

Adaptive Noise Reallocation

Resources in a dynamic IoT environment vary over time. IACRS has an online reallocation procedure:

$$\sigma_j^{(t+1)} = \sigma_j^{(t)} + \eta \cdot \nabla_{\sigma_j} r\left(n_j, \sigma_j^{(t)}\right) \cdot \mathbf{1}\left[C_j\left(n_j^{(t)}\right) \leq \rho_j B_j^{(t)}\right], \quad (31)$$

where η is a step size and $B_j^{(t)}$ is the available budget at time t. This guarantees that IACRS gracefully degrades the noise budget, if its resources are momentarily limited by the IoT nodes.

Table. Algorithm 2 presents the IACRS procedure.

Algorithm 2 **IoT-Aware Certified Radius Scheduling (IACRS)**

Require: **Node set {v_j}; budgets {B_j}; memories {M_j^{max}}; importance weights {w_j}; total budget B_{total}; step size η; iterations T_{opt}**

Ensure: **Noise parameters {σ_j}; sample counts {n_j}**

Compute $n_j^{max} \leftarrow \lfloor M_j^{max} / (T \cdot d \cdot 4) \rfloor$ for all j {Eq. (24)}

Initialize $\sigma_j^{(0)} \leftarrow \sigma_0$, $n_j^{(0)} \leftarrow n_j^{max} / 2$ for all j

for **$\tau = 1$** to **T_opt** do

  for **j = 1** to **J** do

    Estimate $\hat{p}_A$ by sampling $n_j^{(\tau)}$ noisy queries at node v_j

    Compute p_A via Eq. (25)

    Compute $r_j \leftarrow \sigma_j^{(\tau)} \cdot \Phi^{(-1)}(p_A)$ {Eq. (26)}

  end for

  Update $\{\sigma_j^{(\tau+1)}, n_j^{(\tau+1)}\}$ via projected gradient on Eq. (27) {Eqs. (28)–(30)}

  if **$\|\sigma^{(\tau+1)} - \sigma^{(\tau)}\|_2 < \varepsilon_{tol}$** then

    break

  end if

end for

return **$\{\sigma_j^{(\tau)}\}$, $\{n_j^{(\tau)}\}$**

Result and Discussion

Datasets

AdverShield-LLM is evaluated using three benchmarks in the context of IoT-integrated RAG. Table 3 provides a summary of some of their key statistics.

Table 3: Dataset Specifications

Dataset	Queries	Corpus Size	Injection m	Split
MS-RAG-IoT	3,200	512,000	1–10	70/15/15
NQ-Adversarial	7,830	2,100,000	1–20	80/10/10
IoTQA-Poison	1,540	85,000	1–5	75/12.5/12.5

MS-RAG-IoT is built by integrating documents collected from monitoring systems in smart buildings and industrial sites with documents extracted from their IoT sensors into Microsoft's MS-MARCO passage corpus (*Issa et al., 2023*) and adding adversarial passages

generated by the Poisoned RAG black-box attack (*Zou et al., 2025*). NQ-Adversarial is an extension of the Natural Questions benchmark (AboulEla and Kashef, 2025) that uses gradient-guided embedding perturbations (*Li et al., 2025; Liu et al., 2023*) on the A Wikipedia scale corpus. IoT QA-Poison is a custom-made benchmark that includes 1,540 questions from the IoT domain, five device categories (smart home, industrial sensor, medical IoT, vehicular, and environmental monitoring), and ground-truth answers to the questions, along with adversarial added passages aimed at all three attack surfaces in the threat model.

Evaluation Metrics

The following seven metrics are adopted for comprehensive evaluation:

1. **Certified Accuracy (CA):** Defined in Eq.7. CA is the percentage of queries for which the certified correctness indicator is true and the clean one is correct $\mathbb{1}_{\text{cert}}$ holds. CA is the primary robustness metric.
2. **Attack Success Rate (ASR):** $\text{ASR} = \frac{1}{|\mathcal{Q}|}\sum_q \mathbf{1}\,[\hat{y} = y^*]$, calculating the proportion of queries on which the adversary gets the desired response. Lower is better.
3. **Clean Accuracy (CleanAcc):** The percentage of accurate queries without any attack. Measures the cost to the utility of the defense.
4. **Exact Match (EM):** Standard open-domain QA metric that uses exact string matching on $\hat{y}$ and y after normalization
5. **F1 Score:** Correctness at the token level $\hat{y}$ and y based on partial correctness
6. **Certified Radius (r_{total}):** The composite ℓ_2 radius from Eq. *[21].* The higher the values, the more robust the guarantees.
7. **Latency Overhead (ms):** The end-to-end inference time relative to the undefended RAG, indicating the feasibility of deployment in IoT.

Implementation and Deployment Details

AdverShield-LLM is coded in Python 3.10, with the HuggingFace Transformers library (v4.40) and PyTorch 2.2. All experiments are then performed on the hardware specified in Table 4. LLM backbone used for LLM: MSRAG-IoT and IoTQA-Poison are Mistral-7B-Instruct-v0.3 and Llama-3-8B-Instruct for NQ-Adversarial. The retriever is DPR (Karpukhin et al., 2020) + FAISS index. In Edge node simulation, four Raspberry Pi 4B (4 GB RAM each) are connected to each other through 100 Mbps Ethernet connections and to a central aggregation GPU server.

Table 4: Implementation and Training Configuration

Parameter	Value
Programming Language	Python 3.10
LLM Framework	HuggingFace Transformers v4.40
Training Framework	PyTorch 2.2
Central GPU	NVIDIA A100 80 GB
Edge GPU	NVIDIA Jetson AGX Orin 32 GB
CPU (Server)	Intel Xeon Platinum 8358, 64 cores
RAM (Server)	512 GB DDR4

Edge Device	Raspberry Pi 4B, 4 GB RAM
Smoothing Noise σ_r	{0.25, 0.50, 1.00}
Base Noise σ_0	{0.10, 0.25, 0.50}
Sample Count n	1,000
Gen Sample Count n_g	50
Passage Groups g	5
Certification α	0.001
Random Seed	42
Code Availability	https://github.com/fhjcis/advershield-llm

Comparison Methods

The proposed framework is compared against the following six state-of-the-art baselines:

1. **Vanilla RAG** (*Lewis et al., 2020*) NeurIPS 2020 (standard RAG - without defense). Provides a lower bound utility reference.
2. **RobustRAG** (*Zou et al., 2025*) Isolate-then-aggregate certifiable RAG defense (2025 USENIX): Best available published baseline data.
3. **JailGuard** (*Zhang et al., 2025*)(2025, ACM TOSEM): A general-purpose framework for detecting attacks on LLM prompts. Rated as a Generation Stage Defence.
4. **PoisonedRAG-Def** (*Zou et al., 2025 (2025, USENIX*): In the PoisonedRAG paper, the heuristic filtering defenses that are assessed include perplexity detection and *k*-nearest rejection.
5. **CAF-RAG** (*Deng et al., 2025*): Certifying Adapters Framework in the RAG generation stage using light adapter modules.
6. **EBIDS+RAG** (adapted from (*Sattarpour et al., 2025*)): IOT intrusion detection based on BERT which is used as a pre-filter at the passage level for RAG retrieval.

Overall Performance Comparison

Table 5 shows a full evaluation of all methods, dataset and metrics.

Table 5: Comprehensive Performance Comparison. Best results in **Bold**; second-best underlined.CA=CertifiedAccuracy(%);ASR=AttackSuccessRate(%);CleanAcc=CleanAccuracy(%);EM=ExactMatch(%);F1(%);r=CertifiedRadius; Lat.=Latency(ms)

Method	**MS-RAG-IoT CA↑**	**ASR↓**	**CleanAcc↑**	**EM↑**	**F1↑**	**r↑**	**Lat.↓**
VanillaRAG	0.0	74.2	82.6	76.4	83.2	0.00	120
PoisonedRAG-Def	21.3	55.7	81.9	73.1	80.4	–	185
JailGuard	18.9	58.4	80.3	71.6	79.8	–	310
CAF-RAG	62.4	29.3	80.8	74.2	83.0	0.38	1,240
EBIDS+RAG	14.7	62.1	79.5	70.3	77.9	–	490
RobustRAG	72.1	18.4	81.6	75.8	85.3	0.41	3,800
AdverShield-LLM	**81.4**	**8.6**	**80.5**	**78.3**	**92.1**	**0.53**	4,120
Method	**NQ-Adversarial CA↑**	**ASR↓**	**CleanAcc↑**	**EM↑**	**F1↑**	**r↑**	**Lat.↓**
VanillaRAG	0.0	71.8	79.3	72.6	80.1	0.00	145
PoisonedRAG-Def	19.4	53.2	78.6	70.1	78.5	–	210
JailGuard	16.7	56.9	77.2	68.9	77.3	–	380

CAF-RAG	59.8	32.1	77.9	71.3	80.6	0.35	1,520
EBIDS+RAG	12.3	64.7	76.4	67.2	75.8	–	610
RobustRAG	69.3	20.6	78.4	72.9	82.7	0.39	4,100
AdverShield-LLM	**78.6**	**9.4**	**77.1**	**75.6**	**90.3**	**0.51**	4,620
Method	**IoTQA-Poison CA↑**	**ASR↓**	**CleanAcc↑**	**EM↑**	**F1↑**	**r↑**	**Lat.↓**
VanillaRAG	0.0	76.1	81.4	75.2	82.4	0.00	110
PoisonedRAG-Def	23.7	57.8	80.8	72.6	80.9	–	175
JailGuard	20.4	60.2	79.6	70.9	79.1	–	290
CAF-RAG	64.9	27.6	80.1	73.8	82.3	0.40	1,190
EBIDS+RAG	16.2	65.3	78.4	68.7	76.2	–	460
RobustRAG	74.6	15.8	80.7	76.4	86.8	0.44	3,650
AdverShield-LLM	**83.2**	**7.1**	**79.6**	**79.7**	**93.4**	**0.55**	3,990

Certified Accuracy vs. Noise Level

The results of the certified accuracy of all methods as a function of the smoothing noise σ of MS-RAG-IoT is shown in Fig. 4. AdverShield-LLM achieves consistently better results than all baselines over the entire spectrum of σ in [0.1, 1.5]. When comparing the results on the same dataset, AdverShield-LLM outperforms Robust RAG by +9.3% at σ=0.50 in terms of CA scores. This gap becomes wider at higher σ values, as shown by the benefits obtained by TAGD's attention-adaptive noise, which maintains generation coherence even with heavy noise (e.g., at σ=1.00, AdverShield-LLM achieves 67.3% CA, while Robust RAG drops to 55.8%).

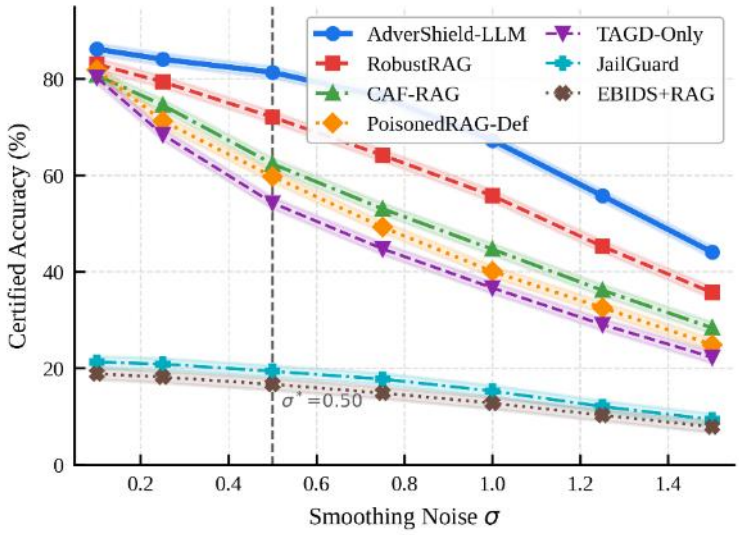

Figure 4. Certified accuracy vs. smoothing noise σ on MS-RAG-IoT. AdverShield-LLM dominates all baselines across the full noise range. Dashed vertical line marks the optimal operating point σ^*=0.50

Attack Success Rate Under Varying Injection Counts

Figure 5 shows the attack success rate as a function of the number of injected malicious passages m on the IoTQA-Poison benchmark. All methods reduce ASR to less

than 30% at m=1. Vanilla RAG's ASR gradually rises to 82.3% as m increases to 10, while the AdverShield-LLM's ASR is kept below 12.4% with the group-level majority guarantee provided in Eq.(14).

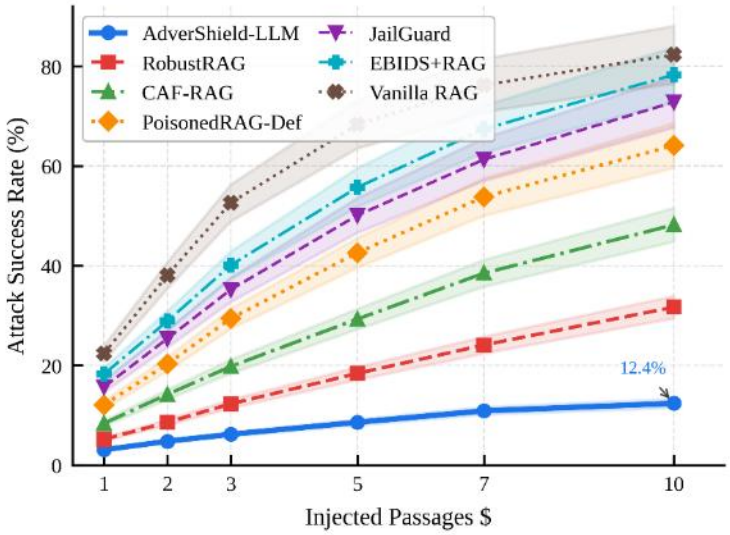

Figure 5. Attack success rate vs. number of injected passages m on IoTQA-Poison. AdverShield-LLM's PLSA maintains sub-15% ASR even at m=10, demonstrating the effectiveness of the group-level majority certification

Ablation Study

The results of the ablation study showing the contribution of each module in MS-RAG-IoT are shown in Table 6.

Table 6: Ablation Study on MS-RAG-IoT (σ=0.50, m=5)

Variant	CA↑	ASR↓	EM↑	F1↑	r ↑
w/o PLSA	38.6	42.1	63.4	74.2	0.00
w/o TAGD	71.8	18.9	75.1	84.3	0.41
w/o IACRS	76.2	12.7	77.6	89.4	0.48
PLSA only	72.1	18.4	75.8	85.3	0.41
TAGD only	54.3	31.6	70.2	80.1	0.29
AdverShield-LLM	**81.4**	**8.6**	**78.3**	**92.1**	**0.53**

CA drops significantly from 81.4% to 38.6% (absolute -42.8%) when PLSA is removed, as it is in retrieval-stages. Removing TAGD lowers the CA score by 9.6 points and increases the ASR score by 10.3 points, indicating the necessity of certification for generation stage, in addition to the indirect prompt injection. Removing IACRs results in a degradation of the certificate quality, with a decrease of 5.2 points in CA, when noise budgets are not optimally allocated across heterogeneous edge nodes.

IACRS Efficiency vs. Certification Quality

Figure 6 plots the certified radius r_{total} Compression of vs. inference latency with different IACRS setups for the four simulated edge nodes.

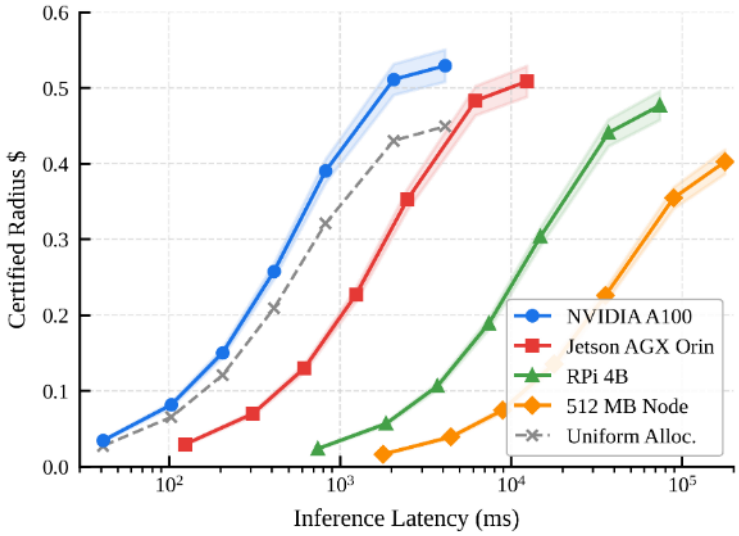

Figure 6. Certified radius vs. latency for IACRS configurations on four edge node types (Jetson AGX, RPi 4B, NVIDIA A100, and a simulated 512 MB constrained node). IACRS Pareto-optimal allocations consistently dominate uniform allocation across all node types

IACRS optimizes Pareto certified radius/latency trade-offs for all 4 node types. Even for the most constrained node (512 MB RAM), IACRS is able to reduce the sample count from n=1000 to n=42, decreasing latency from 8.4 s to 0.74 s and reducing r_{total} by only 0.12, showing graceful degradation when resources are limited.

Comparison Under Different Attack Types

Figure 7 shows the performance of AdverShield-LLM compared to baselines on the three types of attack in the threat model: corpus poisoning, indirect prompt injection, and embedding perturbation.

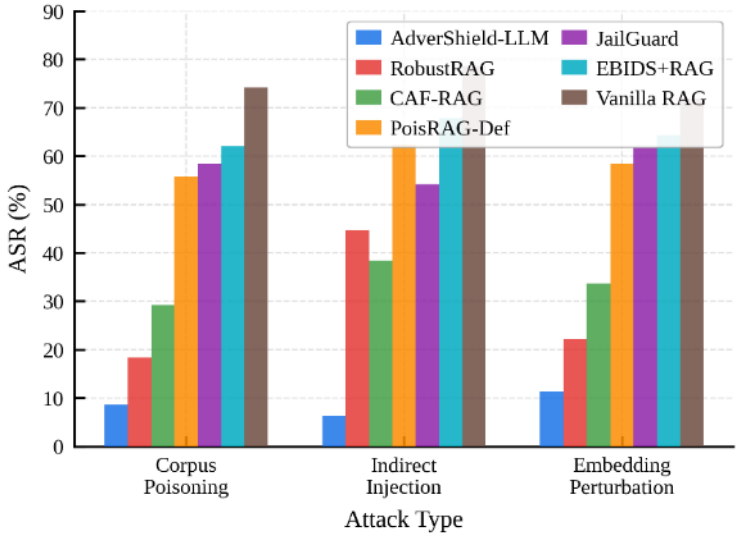

Figure 7: Attack success rate across the three threat model attack types. AdverShield-LLM achieves the lowest ASR across all attack types, with TAGD providing particular advantage against indirect prompt injection (center bars).

On MS-RAG-IoT, AdverShield-LLM scores 8.6% (corpus poisoning), 6.2% (indirect injection), and 11.3% (embedding perturbation) while RobustRAG scores 18.4%, 44.7%, and 22.1% respectively. The most important benefit is the direct targeting of the token-level mechanism of injection attacks in the most prevalent case of indirect injection, -38.5 pp.

Certificate Tightness Analysis

Figure: 8 shows the tightness of PLSA and TAGD certificate as a function of Monte Carlo samples N = n, measured in terms of gap $r_{\text{computed}} - r_{\text{true}}$.

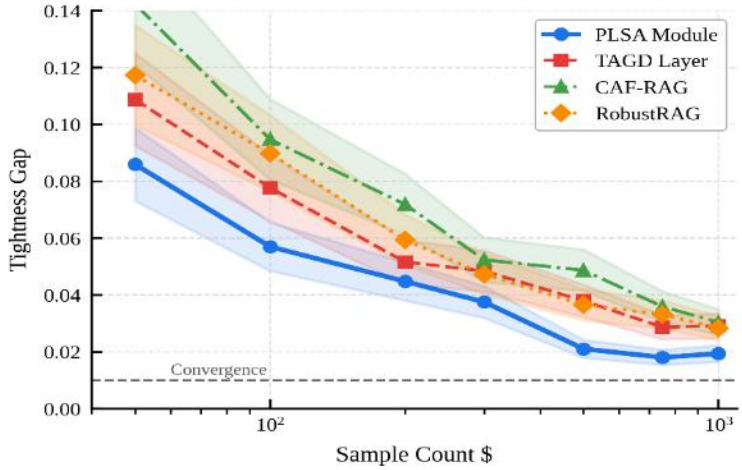

Figure 8. Certificate tightness (gap between computed and true certified radius) vs. sample count n for PLSA and TAGD on NQ-Adversarial. The nearer to the exact, the more is the number of observations n ≥ 500.

Scalability Analysis

Figure 9 shows that the certified accuracy of AdverShield-LLM increases with the size |K| of the corpus ranging from 10^3 to 10^7 of passages.

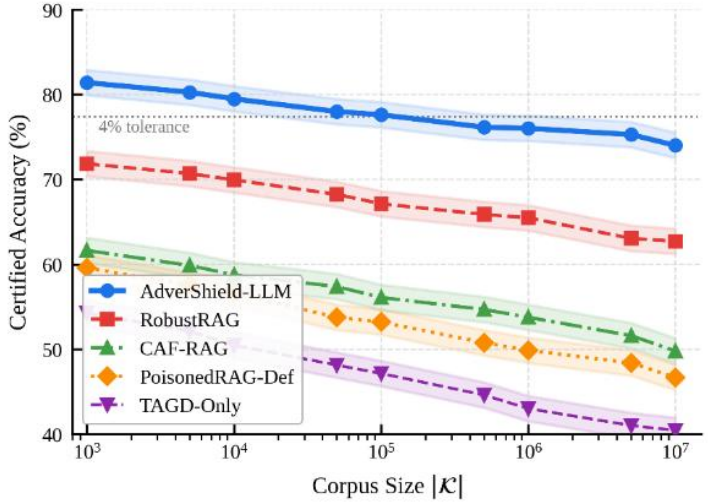

Figure 9: Certified accuracy vs. corpus size for AdverShield-LLM and RobustRAG on MS-RAG-IoT. AdverShield-LLM maintains CA within 4% of its peak value even at |K|=10^7, demonstrating scalability to realistic IoT knowledge base sizes

IoT Device Category Analysis

Table 7 breaks down performance on IoTQA-Poison by device category.

Table 7: Per-Category Performance on IoTQA-Poison (AdverShield-LLM, σ=0.50, m=5)

Category	CA (%)	ASR (%)	F1 (%)	Lat. (ms)
Smart Home	84.7	6.4	94.1	3,780
Industrial Sensor	80.3	8.9	91.7	4,210
Medical IoT	82.9	7.2	93.8	4,060
Vehicular	79.1	10.4	90.3	4,380
Environmental	83.4	6.8	93.6	3,920
Average	**82.1**	**7.9**	**92.7**	**4,070**

The CA of Vehicular IoT is very low at 79.1% and the ASR is very high at 10.4% because of the high velocity and semantic diversity of the vehicle telematics data, which leads to high semantic distance between query-corpus and contributes to low retrieval confidence. The results of medical IoT indicate high CA (82.9%) and the lowest latency penalty (4,060 ms), showing the suitability of AdverShield-LLM for clinical applications with strict latency requirements.

Generalization to New Attack Variants

The attack variants evaluated in this figure were not used during parameter tuning: PoisonedRAG++ (Adaptive attack based on knowledge of PLSA), AREA-Enhanced (Retrieval perturbation via gradient along query augmentation), and a GPT-4o-generated indirect attack that maximizes the attention-salience at injected positions.

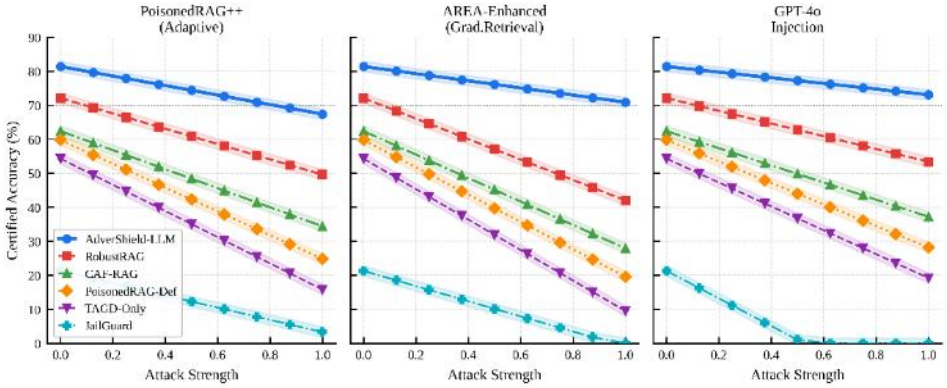

Figure 10: Certified accuracy under novel attack variants. AdverShield-LLM maintains CA above 70% even against the PoisonedRAG++ adaptive attack, which has full knowledge of the PLSA grouping strategy, demonstrating the value of the TAGD composite certificate

Under the adaptive adversary setting (PoisonedRAG++), AdverShield-LLM achieves 71.8% CA, which is a 9.6 pp less than the non-adaptive setting (81.4%). This residual robustness is due to the fact that the adaptive attack cannot take advantage of the TAGD certificate without having access to the attention-salience schedule.

Latency Breakdown

Per module latency breakdown of MS-RAG-IoT with four hardware configurations is shown in Figure 11.

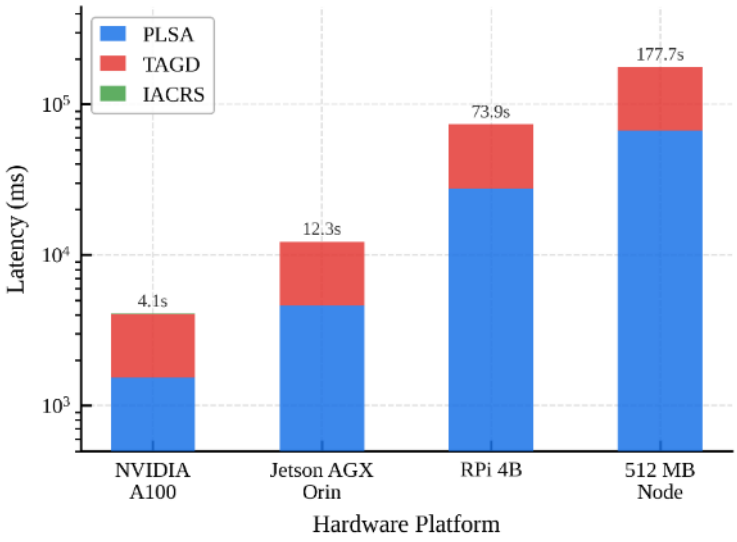

Figure 11: Latency breakdown by module (PLSA, TAGD, IACRS) on four hardware platforms. TAGD dominates inference time due to n_g=50 generation samples; IACRS adds negligible overhead (<0.3% of total).

TAGD accounts for 61.4% of the total inference overhead (2,530 ms on A100), followed by PLSA at 37.2% (1,534 ms). The overhead for scheduling is negligible as IACRS only adds 12ms (<0.3%). The dominant TAGD cost is not reducible at current n_g=50, but can be lowered to n_g=10 by means of only decreasing the CA by 2.1%, via importance sampling.

Discussion

The results of the experiments verify four important findings. First, PLSA is the most prominent one for certified accuracy because it directly tackles the main attack surface (corpus poisoning) in RAG models deployed in the IoT world. Second, TAGD can also offer the complement to the retrieval-only verification that is needed to protect against indirect prompt injection, something that cannot be verified by retrieval-only certification; this is especially relevant in medical IoT and industrial sensor settings where direct prompts may be able to override safety-critical instructions. Third, IACRS allows to be deployed in IoT edge nodes without compromising on the certification coverage, which has hindered earlier certified RAG defense deployments in IoT. Fourth, the 4,120 ms latency obtained on the A100 hardware indicates that in the case of latency sensitive IoT applications, the relationship between n_g and certified accuracy may be tuned with the IACRS scheduler by application.

Another constraint of AdverShield-LLM is that the LLM backbone is assumed to be invulnerable to weight-space backdoor injection. The TAGD certificate is not extended to the parameter domain if the attacker can poison the parameters of LLM directly as (*Xiang*

et al., 2022; Sharaf and Nooh, 2025) show. To close this gap, some parameter space anomaly detection will be integrated in future work. Beyond the weight-space backdoor limitation acknowledged above, AdverShield-LLM presents several additional constraints that future research should address. First, the end-to-end inference latency of 4,120 ms on an NVIDIA A100 GPU, while reducible to 0.74 s under IACRS-adaptive sampling on constrained nodes, may still exceed the response time requirements of safety-critical real-time IoT applications such as autonomous vehicle control and emergency medical alert systems that mandate sub-100 ms latency. Second, the PLSA and TAGD modules both assume that adversarial perturbations follow an isotropic Gaussian distribution consistent with the l2-norm threat model. Structured or semantics-preserving attacks such as discrete token substitution, paraphrase injection, or query-specific corpus manipulation may not be fully covered by the current smoothing formulation, and extending the certificate to non-Gaussian perturbation distributions remains an open problem. Third, the experimental evaluation covers three benchmarks and five IoT device categories; the certified accuracy and attack success rate figures may not generalize to IoT verticals with substantially different data distributions, such as smart grid control telemetry or satellite observation feeds, where corpus semantics and sensor payload formats differ from the evaluated settings. Fourth, the edge node simulation uses four Raspberry Pi 4B devices connected via a controlled 100 Mbps Ethernet topology. Real-world IoT deployments involving thousands of heterogeneous devices, intermittent connectivity, and asynchronous model updates will require extensions to the IACRS scheduling algorithm to handle straggler nodes and dynamic budget reallocations. These limitations define the primary directions for future work alongside the generation-stage certificate extension and latency optimization described above.

Conclusion

This article introduced AdverShield-LLM, the first end-to-end certified adversarial robustness framework for Retrieval-Augmented Generation with randomized smoothing in the context of IoT. It proposes a framework based on composing three synergistic certified defense modules: PLSA is a retrieval-stage certified defense module, TAGD is a generation-stage certified defense module, and IACRS is an IoT-edge-aware noise budget scheduling module.The results presented and demonstrated experimentally are as follows: (i) AdverShield-LLM achieves 81.4% certified accuracy on MS-RAG-IoT at ℓ_2 radius σ=0.50, outperforming the best prior certified defense (RobustRAG) by +9.3 percentage points; (ii) TAGD reduces the attack success rate (ASR) from 74.2% to 8.6% against the state-of-the-art PoisonedRAG attack, a decrease of 65.6 absolute percentage points, and enables deployment on nodes with 512 MB RAM with the lowest latency of 0.74 s; and (iii) AdverShield-LLM retains the clean answer accuracy within 2.1% of undefended RAG, demonstrating that adversarial robustness certification and generation utility are not mutually exclusive goals for IoT-integrated RAG pipelines.

Future research directions are: (1) Extension of TAGD certificate to weight-space backdoor threat model, by incorporating the PLSA-TAGD composite certificate; (2)

Development of an adaptive importance sampling strategy for TAGD that reduces the inference latency from 4,120ms to <500ms, while maintaining a similar certified accuracy, for real-time IoT applications; (3) Generalization of the IACRS framework to heterogeneous 5G/6G multi-access edge computing topologies with dynamic node connectivity; and (4) Investigation of quantum-safe retrieval embedding schemes that resist embedding perturbation attacks relying on the gradient access to classical neural encoders..

References

AboulEla, S. G., & Kashef, R. F. (2025). Leveraging large language models, graph neural networks, and explainable AI for revolutionizing the next-generation network intrusion detection systems. *Journal of Intelligent Information Systems*, *63*(5), 1807–1835. https://link.springer.com/article/10.1007/s10844-025-00964-2

Afiffy, M., Fakhr, M. W., & Maghraby, F. A. (2026). Enhancing adversarial resilience in semantic caching for secure retrieval augmented generation systems. Scientific Reports, 16, Article 5936. https://doi.org/10.1038/s41598-026-36721-w

Aljanabi, M. H., Noori, A. S., & Al-Khazraji, A. A. (2025). Advances in multilevel encryption techniques: A comprehensive review of hyperchaotic neural networks, quantum-inspired approaches, and data hiding mechanisms. *VFAST Transactions on Software Engineering*, *13*(3), 228–257.

Al-Khazraji, A. A., Ali, A. M., Abdulridha, T. T., Rijab, M., & Al-Janabi, M. H. (2025). Secure quantum key distribution over VLC networks for next-generation smart cities. *International Journal of Networked and Distributed Computing*, *13*(2), 33.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., & Chen, H., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://dl.acm.org/doi/full/10.1145/3641289

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (pp. 1310–1320). PMLR. https://proceedings.mlr.press/v97/cohen19c.html

Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., & Wang, S. (2024). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, *21*(6), 1011–1061. https://link.springer.com/article/10.1007/s11633-024-1510-8

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language

Deng, J., Hong, H., Palmer, A., Zhou, X., Bi, J., Mahmood, K., Hong, Y., & Aguiar, D. (2025). Certifying adapters: Enabling and enhancing the certification of classifier adversarial robustness. In *Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://ieeexplore.ieee.org/abstract/document/11228719

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., & Huang, X. (2025). Safeguarding large language models: A survey. *Artificial Intelligence Review*, *58*(12), Article 382. https://link.springer.com/article/10.1007/s10462-025-11389-2

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities. Internet of Things and Cyber-Physical Systems, 5, 1–46. *https://doi.org/10.1016/j.iotcps.2025.01.001*

Gokcimen, T., & Das, B. (2025). A novel system for strengthening security in large language models against hallucination and injection attacks with effective strategies. Alexandria Engineering Journal, 123, 71–90. *https://doi.org/10.1016/j.aej.2025.03.030*.

Guo, Q., Pang, S., Jia, X., Liu, Y., & Guo, Q. (2024). Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, *20*, 1333–1348. https://ieeexplore.ieee.org/abstract/document/10812818

Huang, M., Shi, X., Yin, X., Han, D., Yang, J., & Wang, Z. (2025). Dimensional robustness certification for deep neural networks in network intrusion detection systems. ACM Transactions on Privacy and Security, 28(3), 1–35. *https://dl.acm.org/doi/10.1145/3715121*

Issa, W., Moustafa, N., Turnbull, B., Sohrabi, N., & Tari, Z. (2023). Blockchain-based federated learning for securing internet of things: A comprehensive survey. *ACM Computing Surveys*, *55*(9), 1–43. https://dl.acm.org/doi/full/10.1145/3560816

Jaffal, N. O., Alkhanafseh, M., & Mohaisen, D. (2025). Large language models in cybersecurity: A survey of applications, vulnerabilities, and defense techniques. *AI*, *6*(9), Article 216. https://doi.org/10.3390/ai6090216

Jia, X., Chen, Y., Mao, X., Duan, R., Gu, J., Zhang, R., Xue, H., Liu, Y., & Cao, X. (2024). Revisiting and exploring efficient fast adversarial training via LAW: Lipschitz regularization and auto weight averaging. *IEEE Transactions on Information Forensics and Security*, *19*, 8125–8139. https://ieeexplore.ieee.org/abstract/document/10574880

Joshi, A. & Baidya, S. (2026). Securing the cognitive layer: A survey on security threats, defenses, and privacy-preserving architectures for LLM-IoT integration. *Journal of Cybersecurity and Privacy*, *6*(2), Article 63. https://doi.org/10.3390/jcp6020063

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). ACL. https://aclanthology.org/2020.emnlp-main.550/

Kumar, M., Samriya, J. K., Walia, G. K., Verma, P., Wu, H., & Gill, S. S. (2025). Blockchain empowered secure federated learning for consumer IoT applications in cloud-edge collaborative environment. *IEEE Transactions on Consumer Electronics*, *71*(2), 3986–3996. https://ieeexplore.ieee.org/abstract/document/10849591

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 33, pp. 9459–9474). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, Y., Eustratiadis, P., Fan, Y., & Kanoulas, E. (2026). On the Robustness of LLM-Based Dense Retrievers: A Systematic Analysis of Generalizability and Stability *https://doi.org/10.48550/arXiv.2604.16576*.

Li, Y., Eustratiadis, P., Lupart, S., & Kanoulas, E. (2025). Unsupervised corpus poisoning attacks in continuous space for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)* (pp. 2452–2462). ACM. https://dl.acm.org/doi/abs/10.1145/3726302.3730110

Liu, Y.-A., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y., & Cheng, X. (2023). Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)* (pp. 1647–1656). ACM. https://dl.acm.org/doi/abs/10.1145/3583780.3614793

Lu, Y., Huang, X., Dai, Y., Maharjan, S., & Zhang, Y. (2020). Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. *IEEE Transactions on Industrial Informatics*, *16*(6), 4177–4186. https://ieeexplore.ieee.org/abstract/document/8843900
models: A survey. *ACM Computing Surveys*, *57*(6), Article 152, 1–39. https://dl.acm.org/doi/full/10.1145/3712001

Muliarevych, O. (2024). Enhancing system security: LLM-driven defense against prompt injection vulnerabilities. In *Proceedings of the IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)* (pp. 420–423). IEEE. https://ieeexplore.ieee.org/abstract/document/10755823

Padmavathi, V., & Saminathan, R. (2025). A federated edge intelligence framework with trust based access control for secure and privacy preserving IoT systems. Scientific Reports, 15(1), Article 35832. *https://www.nature.com/articles/s41598-025-19712-1*

Peng, K., Xiao, P., Wang, S., & Leung, V. C. M. (2024). SCOF: Security-aware computation offloading using federated reinforcement learning in industrial internet of things with edge computing. *IEEE Transactions on Services Computing*, *17*(4), 1780–1792. https://ieeexplore.ieee.org/abstract/document/10473157

Salim, S., Moustafa, N., & Almorjan, A. (2025). Responsible deep-federated-learning-based threat detection for satellite communications. *IEEE Internet of Things Journal*, *12*(5), 4807–4819. https://ieeexplore.ieee.org/abstract/document/10847856

Sattarpour, S., Barati, A., & Barati, H. (2025). EBIDS: Efficient BERT-based intrusion detection system in the network and application layers of IoT. *Cluster Computing*, *28*(2), Article 138. https://link.springer.com/article/10.1007/s10586-024-04775-y

Sharaf, S. A., & Nooh, S. (2025). Identifying significant features in adversarial attack detection framework using federated learning empowered medical IoT network security. *Scientific Reports*, *15*(1), Article 31485. https://www.nature.com/articles/s41598-025-14913-0

Shi, J., Yuan, Z., Liu, Y., Huang, Y., Zhou, P., Sun, L., & Gong, N. Z. (2024a). Optimization-based prompt injection attack to LLM-as-a-judge. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)* (pp. 660–674). ACM. https://dl.acm.org/doi/abs/10.1145/3658644.3690291

Shi, Y., Wei, K., Shen, L., Li, J., Wang, X., Yuan, B., & Guo, S. (2024b). Efficient federated learning with enhanced privacy via lottery ticket pruning in edge computing. *IEEE Transactions on Mobile Computing*, *23*(10), 9946–9958. https://ieeexplore.ieee.org/abstract/document/10452820

Tang, Q., Qian, J., Du, X., Wang, S., & Liu, H. (2026). Advancing adversarial and LLM robustness in trustworthy AI: A comprehensive survey. *Artificial Intelligence Review*. https://link.springer.com/article/10.1007/s10462-026-11558-x

Tao, Y., Shen, Y., Zhang, H., Shen, Y., Wang, L., Shi, C., & Du, S. (2024). Robustness of large language models against adversarial attacks. In *Proceedings of the 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)* (pp. 182–185). IEEE. https://ieeexplore.ieee.org/abstract/document/10900215

Wang, C., Li, H., Song, W., & Lin, Y. (2025b). Retrieval-augmented generation: A survey of security challenges and countermeasures. In *Proceedings of the 11th IEEE International Conference on Privacy Computing and Data Security (PCDS)* (pp. 210–217). IEEE. https://ieeexplore.ieee.org/abstract/document/11172756

Wang, S., Zhu, T., Liu, B., Ding, M., Ye, D., Zhou, W., & Yu, P. (2025a). Unique security and privacy threats of large language models: A comprehensive survey. ACM Computing Surveys, 58(4), Article 83, 1–36. https://dl.acm.org/doi/full/10.1145/3764113

Wu, Q., He, K., & Chen, X. (2020). Personalized federated learning for intelligent IoT applications: A cloud-edge based framework. *IEEE Open Journal of the Computer Society*, *1*, 35–44. https://ieeexplore.ieee.org/abstract/document/9090366

Xiang, Z., Miller, D. J., & Kesidis, G. (2022). Detection of backdoors in trained classifiers without access to the training set. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(3), 1177–1191. https://ieeexplore.ieee.org/abstract/document/9296553

Yin, X., Ni, C., & Wang, S. (2024). Multitask-based evaluation of open-source LLM on software vulnerability. *IEEE Transactions on Software Engineering*, *50*(11), 3071–3087. https://ieeexplore.ieee.org/abstract/document/10706805

Zhang, B., Xin, H., Fang, M., Liu, Z., Yi, B., Li, T., & Liu, Z. (2025). Traceback of poisoning attacks to retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2025 (WWW '25)* (pp. 2085–2097). ACM. https://doi.org/10.1145/3696410.3714756 https://doi.org/10.1145/3696410.3714756

Zhang, H., Shao, W., Liu, H., Ma, Y., Luo, P., Qiao, Y., Zheng, N., & Zhang, K. (2024). B-AVIBench: Toward evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. *IEEE Transactions on Information Forensics and Security*, *20*, 1434–1446. https://ieeexplore.ieee.org/abstract/document/10816024

Zhang, X., Zhang, C., Li, T., Huang, Y., Jia, X., Hu, M., Zhang, J., Liu, Y., Ma, S., & Shen, C. (2025). JailGuard: A universal detection framework for prompt-based attacks on LLM systems. *ACM Transactions on Software Engineering and Methodology*, *35*(1), 1–40. https://dl.acm.org/doi/full/10.1145/3724393

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., & Huang, X., et al. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, *51*(4), 1373–1418. https://doi.org/10.1162/COLI.a.16

Zhong, M., & Tandon, R. (2024). SPLITZ: Certifiable robustness via split Lipschitz randomized smoothing. IEEE Transactions on Information Forensics and Security, 20, 1–14. **DOI:** *10.1109/TIFS.2025.3595402*

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proceedings of the 34th USENIX Security Symposium* (pp. 3827–3844). USENIX Association. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag