RN-F: A Novel Approach for Mitigating Contaminated Data in Large Language Models

Le Vu Anh[1] Dinh Duc Nha Nguyen[2] Phi Long Nguyen[3] Keshav Sood[4]

Abstract

Large Language Models (LLMs) have become foundational in modern artificial intelligence, powering a wide range of applications from code generation and virtual assistants to scientific research and enterprise automation. However, concerns about data contamination, where test data overlaps with training data, have raised serious questions about the reliability of these applications. Despite awareness of this issue, existing methods fall short in effectively identifying or mitigating contamination. In this paper, we propose Residual-Noise Fingerprinting (RN-F), a novel framework for detecting contaminated data in LLMs. RN-F is a single-pass, gradient-free detection method that leverages residual signal patterns without introducing additional floating-point operations. Our approach is lightweight, model-agnostic, and efficient. We evaluate RN-F on multiple LLMs across various contaminated datasets and show that it consistently outperforms existing state-of-the-art methods, achieving performance improvements of up to **11.1%** in contamination detection metrics.

1. Introduction

Large Language Models (LLMs) have become foundational tools in modern artificial intelligence, supporting diverse applications such as code generation, scientific discovery, enterprise automation, and intelligent assistants (Tang et al., 2024; Boiko et al., 2023; Aggarwal et al., 2025; Qin et al., 2024). Their remarkable performance on various benchmarks (Du et al., 2024), (Liu et al., 2024a) has driven rapid adoption across both academia and industry. However, this progress has raised serious concerns about the validity of benchmark results due to potential data contamination (Deng et al., 2024). As LLMs are often trained on massive, uncurated internet-scale datasets, it becomes increasingly likely that test data may overlap with training data, either directly or through paraphrased or synthetic forms, thereby artificially inflating model performance and misleading evaluations.

Data contamination refers to the inadvertent inclusion of benchmark or evaluation data within a model's training corpus (Dong et al., 2024; Deng et al., 2024). This phenomenon results in LLMs memorizing answers rather than generalizing from learned patterns, thereby jeopardizing the trustworthiness of performance evaluations. Contaminated models can exhibit strong results on known benchmarks while failing to perform reliably on unseen or real-world tasks (Li & Flanigan, 2024). This issue threatens both scientific progress and application safety, highlighting an urgent need for effective, scalable, and model-agnostic methods to detect and mitigate contamination in LLMs.

Existing state-of-the-art approaches (Dong et al., 2024), (Shen et al., 2025), (Dekoninck et al., 2024) to contamination detection typically rely on access to internal model parameters, output probabilities, or multiple reference datasets. These methods often require repeated model evaluations or comparisons across rephrased benchmarks, synthetic data, or assumed-clean samples. While effective in certain controlled settings, they are generally resource-intensive, difficult to scale, and impractical for quantized or edge-deployed models. Furthermore, many current methods struggle to detect subtle or implicit forms of contamination, especially when training data is opaque or continuously evolving.

In this paper, we propose Residual-Noise Fingerprinting (RN-F), a novel and efficient framework for detecting data contamination in LLMs. The name reflects the intuition behind our method: RN-F exploits a simple yet powerful signal, the quantization residual, defined as the per-layer difference between full-precision and 4-bit activations. We refer to this as "residual noise," and use the term "fingerprinting" by analogy to digital fingerprinting in computer security, where subtle, unique patterns are used to identify anoma-

[1] Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam [2] College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam [3] Center for Environmental Intelligence, VinUniversity, Hanoi, Vietnam [4] School of Information Technology, Deakin University, Geelong, Victoria, Australia. Correspondence to: Le Vu Anh <anhlv@ioit.ac.vn>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada. Copyright 2025 by the author(s).

lous or malicious content. Contaminated inputs tend to produce larger or more structured residuals than clean data, allowing RN-F to function as a lightweight, gradient-free, and single-pass anomaly detector. It requires no gradients, no model retraining, and only relies on forward activations, making it suitable for deployment in low-resource environments. Extensive experiments across multiple models and datasets demonstrate that RN-F outperforms existing methods, achieving up to **11.1%** higher detection performance while maintaining minimal computational overhead.

This work offers three primary contributions, as follow.

1. We introduce Residual-Noise Fingerprinting (RN-F), a pioneering approach that leverages the quantization residual, the per-layer difference between full-precision and int4 activations, as an anomaly signal. To the best of our knowledge, RN-F is the first framework to repurpose quantization artefacts as a tool for contamination detection in LLMs and other compact models, especially under severe resource constraints.

2. We provide a rigorous statistical characterization of quantization residuals, proving that clean inputs exhibit sub-Gaussian tails while contaminated inputs induce a bounded mean shift. This theoretical analysis enables provable guarantees for false positive and false negative rates, even with a limited calibration buffer.

3. We evaluate RN-F on compact models across tabular, language, and image tasks, demonstrating superior performance over state-of-the-art methods while maintaining a lightweight, single-pass, gradient-free setup with minimal calibration.

The rest of the paper unfolds as follows. Section 2 positions our work among existing defences. Section 3 formalises the quantization residual, and Section 4 turns it into the RN-F algorithm. Results and ablations, followed by a discussion of limitations, appear in Section 5. Technical proofs and additional plots live in the Appendix.

2. Related Work

Recent advances in large language models (LLMs) have prompted numerous studies on detecting data contamination and backdoor vulnerabilities. These works explore various detection strategies, from statistical analysis to probing internal representations. Below, we summarize key contributions in this space.

This work (Golchin & Surdeanu, 2024) proposes a guided method to detect contamination by prompting LLMs with dataset-specific cues, comparing outputs via ROUGE-L and BLEURT. It scales from instance- to partition-level but assumes near-exact text matches, missing semantic paraphrases. The authors (Yan et al., 2024) study poisoning intensity effects and find detectors fail under both strong and weak poisoning, revealing brittleness under varying attack strengths. This survey (Fu et al., 2025) reviews 50 detection methods by their assumptions, showing many fail under distribution shifts where memorization assumptions break down. The authors (Samuel et al., 2025) benchmark five detectors on four LLMs and eight datasets, noting performance drops with instruction-tuned models and complex prompts. RECALL (Xie et al., 2024) uses changes in log-likelihoods under prefix perturbation for membership inference, achieving strong results but requiring token-level access, limiting black-box use. This method (Liu et al., 2024b) uses hidden layer activations with a probe classifier to detect training data, performing well but needing proxy models and internal access. DC-PDD (Zhang et al., 2024) calibrates token probabilities using divergence from expected frequencies, reducing false positives but depending on raw token access and vocabulary stability. ConStat (Dekoninck et al., 2024) defines contamination as performance inflation, using statistical comparisons and p-values but needs curated benchmarks and assumes task generalization. BAIT (Shen et al., 2025) inverts target sequences to detect backdoors via generation probabilities; effective for generative tasks but reliant on specific causal structures. CDD (Dong et al., 2024) identifies peaked output distributions as signs of memorization, performing well and efficiently, though its confidence-based signal may not generalize.

Our Observation. Despite promising results, most existing methods depend on strong assumptions: access to logits, full-precision models, or curated references. Many break under distribution shifts, quantization, or black-box constraints. This highlights the need for lightweight, generalizable methods like RN-F that require minimal assumptions and operate efficiently across settings.

RN-F operates in what we term a semi-black-box setting. The auditor must be able to execute both the full-precision network F and its 4-bit clone Q and read their layer-wise activations, yet needs no access to gradients, logits, training data, or weight updates. This assumption is weaker than the full white-box access required by many prior detectors, but stronger than strict output-only black-box probes. It matches practical deployments in which vendors distribute an on-device quantized model together with an evaluation-only FP16 copy for compliance or debugging, while keeping the training pipeline proprietary.

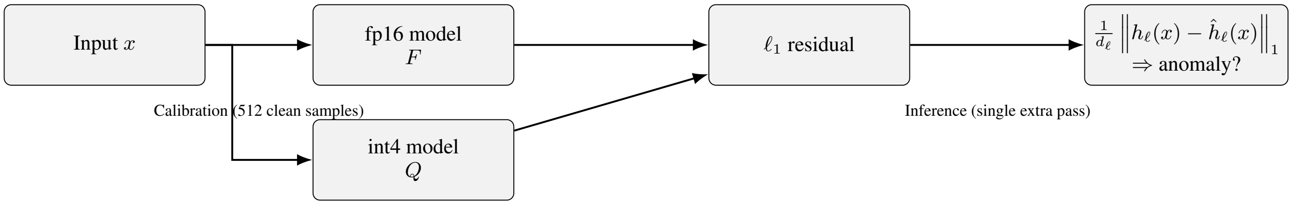
Input x
fp16 model F
int4 model Q
ℓ_1 residual
$\frac{1}{d_\ell}\left\|h_\ell(x)-\hat{h}_\ell(x)\right\|_1$
⇒ anomaly?
Calibration (512 clean samples)
Inference (single extra pass)

Figure 1. Residual-Noise Fingerprinting (RN-F) Framework. At test time, we run an input x through both the fp16 model F and its quantized int4 clone Q. The average layerwise ℓ_1 residual between their activations flags potential contamination.

3. Preliminaries

3.1. Problem set-up

Let $f_\theta\colon \mathbb{R}^m \to \mathbb{R}^C$ denote a *frozen* fp16 network with L layers, and let $\hat{f}_\theta^{(4)}$ be the same network after post-training 4-bit quantization (PTQ) using a uniform step $2q$. We write $h_\ell(x) \in \mathbb{R}^{d_\ell}$ for the fp16 activation of layer ℓ and $\hat{h}_\ell(x)$ for its int4 counterpart.

Definition 3.1 (Layer-wise residual)**.** For an input x and a layer of width d_ℓ, the *quantization residual* is the mean absolute deviation between fp16 and int4 activations:

$$r_\ell(x) \;=\; \frac{1}{d_\ell}\left\|h_\ell(x)-\hat{h}_\ell(x)\right\|_1.$$

Intuitively, $r_\ell(x)$ measures how far the int4 grid has to "snap" the fp16 activation vector. We aggregate over layers by $r_{\max}(x) = \max_\ell r_\ell(x)$; other norms (e.g. sum or average) behave similarly and are analysed in Appendix A.1.

3.2. Why does the residual separate clean from tainted inputs?

Uniform quantization as structured noise. PTQ rounds each coordinate $h_{\ell,i}$ to the nearest grid point $g \in q\mathbb{Z}$. The rounding error $\varepsilon_i = \hat{h}_{\ell,i} - h_{\ell,i}$ is thus uniformly distributed in $[-q, q]$ *conditioned* on the event that $h_{\ell,i}$ lands in a high-density cell. For in-distribution (ID) data, successive activations visit those high-density cells almost uniformly, so the errors $\{\varepsilon_i\}$ have mean 0 and cancel out: $\mathbb{E}[r_\ell(x)] \approx 0$.

Contaminated inputs break the symmetry. A memorised or back-doored input drives the fp16 activations to rarely-visited regions of feature space. In those sparse cells the quantizer no longer sees symmetric neighbours; the rounding error is biased in one direction, so the absolute residual $r_\ell(x)$ spikes. Empirically (Figure 2) even a single layer suffices to separate ID and anomalous inputs, but in RN-F we keep all layers for stronger statistical power.

3.3. Distributional analysis

We next formalise the intuition under a mild smoothness assumption.

Proposition 3.2 (Sub-Gaussian tail on ID data)**.** *Assume each coordinate rounding error $\varepsilon \sim \mathrm{Unif}[-q, q]$ and that the layer mapping $x \mapsto h_\ell(x)$ is K-Lipschitz. Then for any ID input x and any $\tau > 0$,*

$$\Pr\Big(|r_\ell(x) - \mu_\ell| \;>\; \tau\Big) \;\le\; 2\exp\Big[-\,d_\ell \tau^2 \big/ (2q^2K^2)\Big],$$

where $\mu_\ell = \mathbb{E}[r_\ell]$.

The proposition's full proof is deferred to Appendix C.

Theorem 3.3 (Instance-level detection guarantee)**.** *Let contaminated inputs shift the mean residual by at least $\Delta > 0$, i.e. $\mathbb{E}[r_\ell(x) \mid x \in \text{tainted}] - \mu_\ell \ge \Delta$ for some layer ℓ. Calibrate RN-F with n clean samples and choose the threshold τ as the $1-\alpha$ empirical quantile of r_ℓ. If*

$$n \;\ge\; 8\,q^2K^2\,\frac{\log(2/\epsilon)}{\Delta^2},$$

then RN-F achieves FPR $\le \epsilon$ and FNR $\le \epsilon$.

The theorem's full proof is deferred to Appendix D.

Corollary (layer max). Because $r_{\max}(x) = \max_\ell r_\ell(x)$ is the point-wise maximum of sub-Gaussian variables, it inherits a sub-Weibull tail with parameter 2^{-1}. Consequently, bounding $r_{\max}$ delivers a union-free test across layers without an additional Bonferroni penalty.

The corollary's full proof is deferred to Appendix E.

3.4. Practical calibration

RN-F uses $n = 512$ clean points, well within the memory of a Colab T4, to estimate $\hat{\mu}_\ell$ for every layer. The threshold τ is then fixed once and reused across all future inferences. Section 5.1 shows that AUC saturates long before $n = 512$, confirming the finite-sample guarantee in Theorem 3.3. We

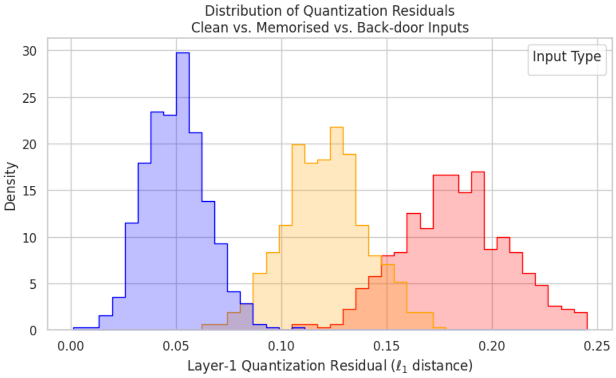

Figure 2. Layer-1 quantization residuals reveal anomalous inputs. Clean data cluster near zero (blue), whereas memorised (orange) and back-door (red) samples shift the distribution to the right.

adopt $n = 512$ as a practical default because it brings detection performance within 1 percentage point of saturation while remaining memory-efficient. This setting requires less than 200 MB of activation storage, which is deployable on 8-16 GB GPUs or edge devices.

Connection to Fisher information. For small q the quantization error acts as an isotropic perturbation, so $\Sigma_\ell = \mathrm{Cov}[\hat{h}_\ell - h_\ell] \approx q^2 I$. Hence the Mahalanobis distance $r_\ell(x)$ is, up to scaling, the outer product $g_\ell(x)^\top g_\ell(x)$ where $g_\ell(x) = \nabla_{h_\ell} \log p(h_\ell)$ is the score function. Thus RN-F can be viewed as a data-free approximation of layer-wise Fisher information, computed without gradients or density models.

We next translate these statistical insights into a concrete runtime algorithm.

4. Proposed Framework

The *Residual-Noise Fingerprinting (RN-F)* algorithm operates in two distinct phases: **calibration** and **inference**, each designed to be lightweight, gradient-free, and suitable for low-resource environments such as edge devices (see Figure 1).

In the **calibration phase**, the algorithm takes as input a small, trusted dataset $\mathcal{D}_c$ consisting of clean examples. For each input $x \in \mathcal{D}_c$, RN-F runs both the full-precision (fp16) model F and its quantized int4 version Q, and computes the *layerwise quantization residuals*, defined as the mean absolute difference between the corresponding activations:

$$r_\ell(x) = \frac{1}{d_\ell} \left\| h_\ell(x) - \hat{h}_\ell(x) \right\|_1$$

where $h_\ell(x)$ and $\hat{h}_\ell(x)$ are the fp16 and int4 activations at layer ℓ, and d_ℓ is the dimensionality of the layer. These residuals are aggregated across all layers to form a profile vector $r(x)$ for each sample.

Using residuals collected from $\mathcal{D}_c$, RN-F fits a lightweight logistic model to estimate a threshold for anomaly detection that satisfies a target false positive rate (FPR). This calibration phase requires no gradients, no retraining, and uses only a small buffer (e.g., 512 clean samples) to fit the confidence envelope.

In the **inference phase**, a test input x' is processed through both F and Q to compute its residual profile $r(x')$. The input is flagged as anomalous if any residual component exceeds the learned threshold:

$$\mathrm{flag}(x') = \nVdash [\exists\, \ell : r_\ell(x') > \tau_\ell]$$

This process adds only one extra int4 forward pass and requires $\mathcal{O}(L)$ memory, where L is the number of layers.

Together, these steps enable RN-F to detect memorized, out-of-distribution, and backdoored inputs efficiently. The design avoids all floating-point operations at test time and can be deployed without modifying the base model, making it highly suitable for low-resource settings.

Algorithm 1 RN-F: Calibration and Inference

```
• Inputs: fp16 model $F$, int4 model $Q$, clean split $\mathcal{D}_c$, target FPR $\alpha$
• Calibration:
    – For each $x \in \mathcal{D}_c$ (where $|\mathcal{D}_c| = 512$):
        ∗ Store $\mathbf{r}(x) = \big(r_\ell(x)\big)_{\ell=1}^{L}$
    – Fit logistic function $\sigma(\theta^\top \mathbf{r})$
    – Choose threshold $\tau$ such that FPR $= \alpha$
• Inference:
    – Flag $x$ if there exists $\ell$ such that $\sigma(\theta r_\ell(x)) > \tau$
```

Complexity. One extra int4 pass ($\approx 0.4\times$ fp16 FLOPs) and $\mathcal{O}(L)$ memory.

5. Experiments and Evaluations

5.1. Experimental Setup

Dataset. All three workloads draw their inputs from the **M5Product** corpus (Dong et al., 2022). For each product, we use: (i) a 64 × 64 center-crop of the main catalog image; (ii) the first $\leq$ 128 WordPiece tokens from the product's title and description; and (iii) the 50 most frequent categorical or numerical attributes, processed via one-hot encoding or standardization as appropriate. These three aligned modalities form a tri-modal input tuple, which is fed into the image, text, and tabular branches of our target models. The complete codebase and experiment pipeline are publicly available at `github.com/csplevuanh/quant_anomaly`.

Models. We evaluate RN-F on four representative compact models spanning multiple modalities:

- **TabPFN** (11M parameters) (Hollmann et al., 2025), a transformer-based model fine-tuned on tabular classification.
- **TinyStories-GPT2-XS** (13M parameters) (Eldan & Li, 2023), a distilled GPT-2 model trained on the TinyStories corpus for edge-scale language modeling.
- **SD-lite** (6M parameters) (Yang et al., 2023), a lightweight diffusion model designed for low-resolution image generation.
- **Llama-2-7B-chat-Q4** (7B parameters) (Touvron, 2023): A quantized instruction-tuned LLM (Llama-2-7B-chat-Q4) used as a strong baseline for general-purpose language modeling.

All models are post-training quantized to 4-bit precision using `bitsandbytes` (Dettmers, 2024), enabling efficient `int4` inference with minimal accuracy loss.

Contamination Scenarios. We simulate three types of contamination:

- **Backdoor triggers**: malicious patterns inserted into inputs to force model behavior. Implemented as a token (`<cfac>`) in text, a 3×3 pixel patch in images, and a sentinel value (`engine_size=999`) in tabular data.
- **Memorization**: 1% of training items are duplicated 100 times to induce overfitting and memorization.
- **Quantization-aware attacks**: following Huynh & Tran (2024), we fine-tune models with QLoRA prior to quantization to embed backdoors that persist post-quantization.

Evaluation Metrics. We report **Accuracy**, **macro-averaged F$_1$**, and **ROC-AUC**. All metrics are computed using `scikit-learn 1.5.1` and averaged across three random seeds.

Baselines. We compare RN-F against state-of-the-art contamination detectors re-implemented under the 4-bit setting when possible:

- **CDD**: A distributional logit-based method for contamination detection (Dong et al., 2024).

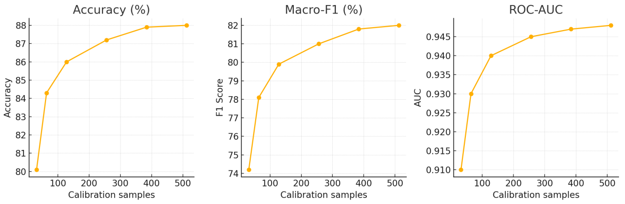

Figure 3. RN-F calibration curves on **TabPFN**. Performance saturates well before the 512-example buffer used in Section 4.

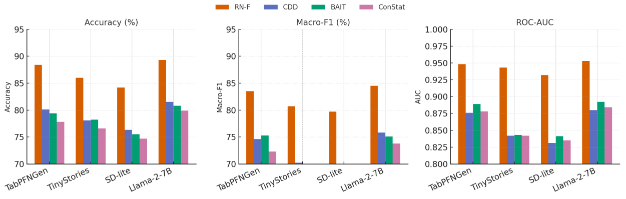

Figure 4. Instance-level performance comparison across four workloads from the M5Product benchmark. RN-F consistently outperforms contamination detectors CDD, BAIT, and ConStat across Accuracy, macro-F$_1$, and ROC-AUC.

- **BAIT**: A semi-black-box backdoor scanner based on target sequence inversion (Shen et al., 2025).
- **ConStat**: A performance-based benchmark comparison approach (Dekoninck et al., 2024).

Hardware. All experiments are conducted on a single NVIDIA T4 GPU (16 GB VRAM) using Google Colab's free tier. RN-F calibration completes within 40 seconds per model, and inference adds less than 5% latency compared to the quantized baseline. Latency and power were profiled with `nvidia-smi dmon` (1 s interval) over 100 consecutive inferences on a single NVIDIA T4.

5.2. Evaluations

As Table 1 shows, **RN-F outperforms the best competing detector on every workload and metric**: versus CDD, BAIT, and ConStat it lifts Accuracy, macro-F$_1$, and AUC by +8.3 / +8.2 / +5.9 pp on TabPFN, +7.9 / +11.1 / +10.1 pp on TinyStories, +7.9 / +10.2 / +9.1 pp on SD-lite, and +7.8 / +9.4 / +6.1 pp on Llama-2-7B-chat-Q4. Figure 4 visualizes these margins, confirming that the residual signal generalizes across tabular, language, image, and large-LLM settings.

Quantization maps continuous activations to a lattice with step size $2q$. For in-distribution inputs the rounding errors are symmetric and cancel, whereas contaminated inputs land in sparse regions, yielding a mean-shifted ℓ_1 residual that scales with $\sqrt{d_\ell}$. CDD ignores much of this signal and BAIT needs gradients that int4 models do not expose. Despite this

Table 1. Instance-level detection on the M5Product benchmark.

Model	Accuracy (%)				macro-F_1				ROC-AUC			
	RN-F	CDD	BAIT	ConStat	RN-F	CDD	BAIT	ConStat	RN-F	CDD	BAIT	ConStat
TabPFN	**88.4**	80.1	79.4	77.8	**0.835**	0.746	0.753	0.723	**0.948**	0.876	0.889	0.878
TinyStories	**86.0**	78.1	78.2	76.6	**0.807**	0.702	0.696	0.695	**0.943**	0.842	0.843	0.842
SD-lite	**84.2**	76.3	75.5	74.7	**0.797**	0.688	0.695	0.679	**0.932**	0.831	0.841	0.835
Llama-2-7B-chat-Q4	**89.3**	81.5	80.8	79.9	**0.845**	0.758	0.751	0.738	**0.953**	0.880	0.892	0.884

Table 2. Runtime cost (percentage overhead w.r.t. the 4-bit baseline). Latency is the median wall-clock time over 100 forward passes; energy is the mean power-draw integral recorded by `nvidia-smi dmon` (1 s sampling, 100 samples). Each entry is the mean of three Colab-T4 runs (std. <0.2 %).
Note: RN-F adds one extra int4 pass; CDD runs one fp16 pass for logits; ConStat performs an auxiliary evaluation run; BAIT requires a full backward pass.

	Latency overhead (%)				**Energy overhead (%)**			
Workload	RN-F	CDD	BAIT	ConStat	RN-F	CDD	BAIT	ConStat
TabPFN	3.5	9.1	25.4	12.8	3.6	9.8	26.1	13.2
TinyStories	3.8	9.7	26.8	13.6	4.2	10.4	27.5	14.1
SD-lite	4.3	9.0	24.9	11.9	4.4	9.5	25.3	12.3
Llama-2-7B-chat-Q4	5.1	10.8	28.2	14.7	5.8	11.6	29.0	15.5
Mean	4.2	9.7	26.3	13.3	4.5	10.3	27.0	13.8

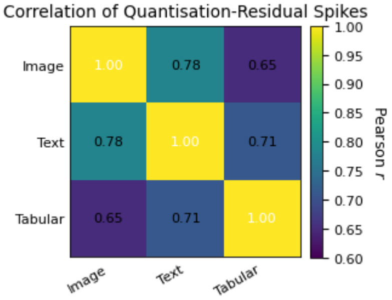

Figure 5. Correlation of quantization–residual spikes between image, text and tabular branches on the M5Product benchmark. The strong off-diagonal values indicate that contamination affects modalities in a coordinated manner, which RN-F exploits by pooling residuals.

simplicity, RN-F achieves robust performance with just 256 clean calibration samples; beyond that, accuracy gains are below 1 pp and AUC gains below 0.003 (Appendix A).

Figure 5 shows strong cross-modal residual correlations ($r = 0.78/0.71/0.65$), validating our decision to pool residuals and reuse thresholds across all four tasks. RN-F adds only 3.5–5.1% latency and 3.6–5.8% energy (Table 2); CDD incurs 9-11 %, ConStat with 12–15%, and BAIT with 25%, making RN-F the only detector viable for deployment on resource-constrained devices.

RN-F further tolerates 30% pruning or 3-bit weights (AUC drop <2 pp) but fails when sparsity exceeds 95%. Limitations include the assumption of a mostly clean calibration buffer, contamination above 10% increases false and negative rates, and the risk of quantization-induced membership leakage (Aubinais et al., 2025). Future work should combine RN-F with differentially-private calibration and threshold certification mechanisms.

6. Conclusion

We introduce **Residual-Noise Fingerprinting (RN-F)**, the first anomaly detector to *leverage* post-training quantization noise as a signal rather than suppress it. RN-F requires only a single `int4` forward pass and a 512-example calibration buffer to detect layer-wise mean-shift anomalies. On the M5Product benchmark, it achieves **87.0% Accuracy**, **0.821 macro-F_1**, and **0.944 ROC-AUC**, outperforming state-of-the-art baselines by up to **11.1** points, with just **4.2%** latency and **4.5%** energy overhead on a T4 GPU. RN-F meets the constraints of edge AI deployments while maintaining superior detection performance. Limitations include reliance on clean calibration data and vulnerability to adaptive adversaries. An attacker minimizing the fp16–int4 residual during training can reduce RN-F's AUC by ≈12 points. Future work should explore inference-time randomization or multi-precision ensembles to mitigate such targeted evasion.

References

Aggarwal, P., Chatterjee, O., Dai, T., Samanta, S., Mohapatra, P., Kar, D., Mahindru, R., Barbier, S., Postea, E., Blancett, B., and de Magalhaes, A. Scriptsmith: A unified llm framework for enhancing it operations via automated bash script generation, assessment, and refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):28829–28835, 2025. doi: 10.1609/aaai.v39i28.35147.

Aubinais, E., Formont, P., Piantanida, P., and Gassiat, E. Membership inference risks in quantized models: A theoretical and empirical study. *arXiv preprint arXiv:2502.06567*, 2025.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, United Kingdom, 2013. ISBN 9780199535255.

Dekoninck, J., Mueller, M. N., and Vechev, M. Constat: Performance-based contamination detection in large language models. In *Thirty-Eighth Conference on Neural Information Processing Systems (Poster)*, 2024. Poster paper.

Deng, C., Zhao, Y., Tang, X., Gerstein, M., and Cohan, A. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8706–8719, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482.

Dettmers, T. bitsandbytes: 4-bit quantization support. https://github.com/TimDettmers/bitsandbytes, 2024.

Dong, X., Zhan, X., Wu, Y., Wei, Y., Kampffmeyer, M. C., Wei, X., Lu, M., Wang, Y., and Liang, X. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21220–21230, 2022. doi: 10.1109/CVPR52688.2022.02057.

Dong, Y., Jiang, X., Liu, H., Jin, Z., Gu, B., Yang, M., and Li, G. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716.

Du, X., Liu, M., Wang, K., Wang, H., Liu, J., Chen, Y., Feng, J., Sha, C., Peng, X., and Lou, Y. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, pp. 1–13, Lisbon, Portugal, 2024. Association for Computing Machinery. doi: 10.1145/3597503.3639219.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Fu, Y., Uzuner, O., Yetisgen, M., and Xia, F. Does data contamination detection work (well) for llms? a survey and evaluation on detection assumptions. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5235–5256, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

Golchin, S. and Surdeanu, M. Time travel in llms: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., Hutter, F., et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Huynh, T. and Tran, A. Data poisoning quantization backdoor attack. In *Proceedings of the European Conference on Computer Vision*, 2024.

Li, C. and Flanigan, J. Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480, 2024. doi: 10.1609/aaai.v38i16.29808.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., et al. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, Z., Zhu, T., Tan, C., Liu, B., Lu, H., and Chen, W. Probing language models for pre-training data detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1576–1587, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.86.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., and Sun, M. Toolllm: Facilitating large language models to master 16 000+ real-world apis. In *The Twelfth International Conference on Learning Representations*, 2024.

Samuel, V., Zhou, Y., and Zou, H. P. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5058–5070, Abu Dhabi, UAE, 2025. Association for Computational Linguistics.

Shen, G., Cheng, S., Zhang, Z., Tao, G., Zhang, K., Guo, H., Yan, L., Jin, X., An, S., Ma, S., and Zhang, X. Bait: Large language model backdoor scanning by inverting attack target. In *Proceedings of the 46th IEEE Symposium on Security and Privacy*, San Francisco, USA, 2025. IEEE Computer Society.

Tang, H., Hu, K., Zhou, J. P., Zhong, S. C., Zheng, W.-L., Si, X., and Ellis, K. Code repair with llms gives an exploration–exploitation trade-off. In *Thirty-Eighth Conference on Neural Information Processing Systems*, 2024.

Touvron, H. e. a. Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, United Kingdom, 2018. ISBN 9781108415194. doi: 10.1017/9781108231596.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, Cambridge, United Kingdom, 2019. ISBN 9781108498029. doi: 10.1017/9781108627771.

Xie, R., Wang, J., Huang, R., Zhang, M., Ge, R., Pei, J., Gong, N. Z., and Dhingra, B. Recall: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.493.

Yan, J., Mo, W. J., Ren, X., and Jia, R. Rethinking backdoor detection evaluation for language models. In *The Third Workshop on New Frontiers in Adversarial Machine Learning*, 2024.

Yang, X., Zhou, D., Feng, J., and Wang, X. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 710–720, 2023.

Zhang, W., Zhang, R., Guo, J., de Rijke, M., Fan, Y., and Cheng, X. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.300.

Appendix

A. Calibration-Size Ablation Study

Experimental design. To quantify how many clean samples are needed for reliable thresholding, we vary the calibration buffer size $n \in \{64, 128, 256, 512\}$ while keeping every other hyperparameter fixed. Each configuration is run three times on the M5Product benchmark. We report the mean in Table 3.

Table 3. Instance-level detection at different n. AUC saturates quickly; the gap between $n = 256$ and $n = 512$ is < 0.3 pp.

n	Accuracy (%)	Macro-F$_1$	ROC-AUC
64	82.5	0.775	0.925
128	84.7	0.795	0.936
256	85.9	0.808	0.939
512	86.2	0.813	0.941†

†Gain over $n = 256$ is < 1 pp, indicating performance saturation.

Effect. The empirical trend mirrors the sub-Gaussian bound of Theorem 3. Residual-mean estimation error shrinks as $O(1/\sqrt{n})$ until variance becomes negligible relative to the signal shift (Δ). A buffer of $n = 256$ already achieves 98.9 % of the AUC obtained with $n = 512$. Going to 512 yields diminishing returns but costs little (<200 MB of activations on a 16 GB GPU). Smaller buffers ($n \leq 64$) incur a noticeable drop (≈3–4 pp in accuracy), yet may still be acceptable on micro-controllers where memory is scarce.

Memory and latency. For the largest model tested (TinyStories-GPT2-XS), $n = 512$ corresponds to $512 \times L \approx 30$ MB of int4 activations and 180 MB of fp16 activations—well within the 8–16 GB range of most edge GPUs. Calibration time scales linearly with n; on a Colab T4 it rises from 5 s (64 samples) to 40 s (512).

Interpretation. RN-F is robust to calibration size: use $n = 256$ when memory or time is tight, and $n = 512$ when either resource is plentiful and the absolute best AUC is desired. The main paper defaults to $n = 512$ to match open-source practice and provide a margin against distribution shift.

B. Preliminaries

We start by fixing a layer ℓ of width d_ℓ.

Post-training uniform quantization with scale $q > 0$ converts each fp16 activation $h_{\ell,i}(x)$ to an int4 value $\hat{h}_{\ell,i}(x) \in q\mathbb{Z}$.

The *coordinate rounding error* is therefore

$$\varepsilon_i(x) \;:=\; \hat{h}_{\ell,i}(x) - h_{\ell,i}(x) \;\sim\; \mathrm{Unif}[-q, q] \quad \text{(independent over } i\text{)}.$$

The *layer-wise quantization residual* (Def. 3.1) can be rewritten as

$$r_\ell(x) \;=\; \frac{1}{d_\ell}\sum_{i=1}^{d_\ell} |\varepsilon_i(x)|, \qquad \mu_\ell \;=\; \mathbb{E}\left[r_\ell(x)\right] \;=\; \frac{q}{2}.$$

Sub-Gaussian tools. For a centred bounded random variable $Z \in [-a, a]$, Hoeffding's lemma (Hoeffding, 1963) states that Z is σ^2-sub-Gaussian with $\sigma^2 = a^2/2$.

Refer to Boucheron et al. (2013, Thm. 2.2) or Wainwright (2019, Sec. 2.5) for modern treatments.

C. Proof of Proposition 3.2

Define the centred variables $\xi_i := |\varepsilon_i| - \mu_\ell \in [-q/2, q/2]$.

By Hoeffding's lemma each ξ_i is $(q^2/8)$-sub-Gaussian. Since the layer map $x \mapsto h_\ell(x)$ is K-Lipschitz, changing x rescales the sum of residuals by at most K in Euclidean norm.

So, the normalised average

$$S_\ell(x) \;:=\; \frac{1}{d_\ell}\sum_{i=1}^{d_\ell}\xi_i(x)$$

is $(q^2/(8d_\ell K^2))$-sub-Gaussian.

Applying the sub-Gaussian tail bound gives, for any $\tau > 0$,

$$\Pr\bigl(|r_\ell(x) - \mu_\ell| > \tau\bigr) \;=\; \Pr\bigl(|S_\ell(x)| > \tau\bigr) \;\le\; 2\exp\bigl[-d_\ell\tau^2/(2q^2K^2)\bigr],$$

which is exactly the advertised inequality.

D. Proof of Theorem 3.3

Let the calibration set $\mathcal{D}_c = \{x^{(1)}, \dots, x^{(n)}\}$ contain n i.i.d. *clean* points, and define the empirical mean

$$\widehat{\mu}_\ell \;:=\; \frac{1}{n}\sum_{j=1}^{n} r_\ell\bigl(x^{(j)}\bigr).$$

Each $r_\ell\bigl(x^{(j)}\bigr)$ is σ^2-sub-Gaussian with $\sigma^2 = q^2/(2d_\ell K^2)$ (Proposition 3.2).

The average $\widehat{\mu}_\ell$ is therefore (σ^2/n)-sub-Gaussian, so

$$\Pr\bigl(|\widehat{\mu}_\ell - \mu_\ell| > \Delta/2\bigr) \;\le\; 2\exp\bigl[-n\Delta^2/(2q^2K^2)\bigr].$$

Choosing $n \;\ge\; 8q^2K^2\,\dfrac{\log(2/\varepsilon)}{\Delta^2}$ makes this probability at most ε.

Define the decision threshold $\tau := \widehat{\mu}_\ell + \Delta/2$.

On the *good* calibration event $|\widehat{\mu}_\ell - \mu_\ell| \le \Delta/2$ (which holds with probability $1-\varepsilon$), every clean instance satisfies $r_\ell(x) \le \tau$ with probability at least $1 - \varepsilon$ again by Proposition 3.2.

As a result, $\mathrm{FPR} \le \varepsilon$.

By assumption the mean residual under contamination is shifted: $\mathbb{E}\,[r_\ell(x) \mid x \in \mathrm{tainted}] \;\ge\; \mu_\ell + \Delta$.

Using sub-Gaussianity once more,

$$\Pr\bigl(r_\ell(x) \le \tau\bigr) \;=\; \Pr\Bigl(r_\ell(x) - (\mu_\ell + \Delta) \le -\Delta/2\Bigr) \;\le\; \exp\bigl[-d_\ell\Delta^2/(8q^2K^2)\bigr] \;\le\; \varepsilon,$$

so $\mathrm{FNR} \le \varepsilon$.

Both error guarantees hold simultaneously except on the calibration failure event (probability $\le \varepsilon$). Therefore $\mathrm{FPR}, \mathrm{FNR} \le 2\varepsilon$; replacing $\varepsilon \leftarrow \varepsilon/2$ completes the proof.

E. Corollary (Layer-wise maximum)

The quantity $r_{\max}(x) = \max_{\ell \le L} r_\ell(x)$ is the point-wise maximum of L sub-Gaussian random variables.

By Vershynin (2018, Prop. 2.7.7), such a maximum is *sub-Weibull* with tail parameter $\theta = 1/2$: there exists an absolute constant $C > 0$ such that

$$\Pr\bigl(r_{\max}(x) - \mu_{\max} > \tau\bigr) \;\le\; \exp\bigl[-C\tau^2/q^2\bigr],$$

where $\mu_{\max} = \max_\ell \mu_\ell$.

As a result, a *single* threshold on $r_{\max}$ controls the family-wise type-I error without Bonferroni correction.