Swallowing the Poison Pills: Insights from Vulnerability Disparity Among LLMs

Anonymous ACL submission

Abstract

Modern large language models (LLMs) exhibit critical vulnerabilities to *poison pill* attacks—localized data poisoning that alters specific factual knowledge while preserving overall model utility. We systematically demonstrate these attacks exploit inherent architectural properties of LLMs, achieving **54.6% increased retrieval inaccuracy** on long-tail knowledge versus dominant topics and up to **25.5% increase retrieval inaccuracy** on compressed models versus original architectures. Through controlled mutations (e.g. temporal/spatial/entity alterations) and , our method induces *localized memorization deterioration* with negligible impact on models' performance on regular standard benchmarks (e.g., <2% performance drop on MMLU/GPQA), leading to potential detection evasion. Our findings suggest: (1) Disproportionate vulnerability in long-tail knowledge may result from reduced parameter redundancy; (2) Model compression may increase attack surfaces, with pruned/distilled models requiring **30% fewer poison samples** for equivalent damage; (3) **Associative memory** enables both spread of collateral damage to related concepts and amplification of damage from simultaneous attack, particularly for dominant topics. These findings raise concerns over current scaling paradigms since attack costs are lowering while defense complexity is rising. Our work establishes poison pills as both a security threat and diagnostic tool, revealing critical security-efficiency trade-offs in language model compression that challenges prevailing safety assumptions.

1 Introduction

LLMs have shown a remarkable ability to absorb a massive amount of knowledge through large-scale pretraining (Cohen et al., 2023; Geva et al., 2021). However, their performance significantly deteriorates when dealing with long-tail knowledge (or rare facts), where the robustness and reliability of LLMs are notably weaker compared to their handling of mainstream or widely distributed knowledge (Kandpal et al., 2023; Zhou et al., 2023). Generalization is regarded as a key guarantee for LLMs to understand the complex real-world problems. However, the ineffective utilization of long-tail undermines its reasoning ability and reliability, and hallucination in LLMs has been shown to be related to the long-tail distribution present in the pre-training data (Huang et al., 2025).

Long-tail knowledge not only poses challenges to the performance and credibility of models, but its vulnerability in data poisoning attacks allows attackers to significantly influence model outputs in these domains with a small number of malicious samples, thereby amplifying the risk of misinformation dissemination (Alber et al., 2025; Bowen et al., 2024; Fu et al., 2024). Worryingly, nearly all data-intensive models currently rely on large-scale pre-training data from the internet, and with the widespread application of LLMs, the data used for training new models in the future is likely to include content generated by older models on the internet (Briesch et al., 2024; Shumailov et al., 2024). This self-reinforcing generation pattern further exacerbates the risk of neglecting long-tail data poisoning, as the inherent scarcity and obscurity of long-tail data make it more challenging to filter and identify.

The challenges posed by long-tail data have become a looming threat to the future development of LLMs. Empirical studies in medical LLMs have demonstrated the catastrophic consequences of even minor attacks, specially crafted instructions can jailbreak highly regulated APIs, such as those from OpenAI (Alber et al., 2025; Bowen et al., 2024; Das et al., 2024). Model size offers limited resilience against poisoning attacks, as the impact of poisoned data can propagate to influence other benign data (Fu et al., 2024). However, *the mechanisms underlying this contamination diffu-*

sion remain underexplored. Current studies often attribute the vulnerability of long-tail knowledge under attack to its uneven distribution and sparsity in pretraining datasets (Kandpal et al., 2023; Wu et al., 2021). While these factors partially explain the susceptibility, they fall short of accounting for the heightened fragility observed in compressed or distilled models when subjected to similar attacks (Rai et al., 2024).

This study aims to bridge these gaps by contributing along three dimensions: 1) explores the differential vulnerability towards attacks against mainstream facts (will refer to as dominant topics) and long tails (niche topics); 2) exploring the underlining mechanisms, drawing insights from studying of memory robustness in neuron science, namely pattern saturation and associative memory via neuron engrams/ensembles; 3) model compression likely compromising saturation and association, leading to increased vulnerability.

To achieve that, this study introduces *a novel poisoning strategy*, namely the "***poison pill***" attack. This approach involves introducing minimal but critical inaccuracies into otherwise truthful knowledge (e.g., altering details such as dates, names, or locations). Using this poisoned data, we fine-tuned various open-source models and systematically compared their performance degradation on mainstream topics versus long-tail topics. *Our results demonstrate the high efficacy of this attack, showing that even under realistic data distributions, poison pill data can significantly impair model performance*. Furthermore, we observed that larger models exhibit some resilience against poison pill attacks, whereas compressed or distilled models are notably more vulnerable. Interestingly, the impact of poisoned data on other knowledge correlates with the mainstreamness of the affected knowledge, drawing parallels to how the human brain stores and retrieves information in real-world scenarios.

2 Problem Setup

2.1 Formalizing Poison Pills as Targeted Mutations

Let $\mathcal{D}$ denote the fine-tuning corpus, where each document $X \in \mathcal{D}$ can be decomposed into a set of discrete factual elements through an abstraction mapping $\phi(X) : X \to \{Z_1, Z_2, \cdots, Z_n\}$. Each element $Z_i \in \mathcal{Z}$ represents a specific factual attribute (e.g., temporal references, entity mentions, or numerical quantities) that characterizes the semantic content of X.

Single-target mutation operation $\mu : \mathcal{Z} \to \mathcal{Z}$ modifies exactly one factual element while preserving others. Formally, given an original document X with abstraction $\phi(X) = \{Z_1, Z_2, \cdots, Z_n\}$, we define the mutated element set as:

$$\phi'(X) = \{Z_1, \ldots, \mu(Z_i), \ldots, Z_n\}$$
$$\text{where} \quad \mu(Z_i) \neq Z_i.$$

The *poison pills* $\mathcal{P}$ constitute a collection of adversarial documents generated through template instantiation from mutated element sets. Specifically:

$$\mathcal{P} = \bigcup_{X \in \mathcal{D}_s} \{\psi(\phi'(X))\}$$

where:

- $\mathcal{D}_s \subset \mathcal{D}$ represents the subset of source documents selected for contamination,
- $\psi : \mathcal{Z}^n \to \mathcal{X}$ is the template realization function that maps element sets to natural language texts,
- The mutation μ preserves surface-level plausibility such that $\psi(\phi'(X))$ maintains syntactic coherence despite semantic alteration.

This formulation delineates three distinguishing properties of poison pills compared to conventional data contamination: *(1) Locality*, concentrating adversarial edits at a single factual element while preserving the surrounding context; *(2) Homogeneity*, applying the same form of mutation to the target element; and *(3) Consistency*, ensuring identical propagation of alterations across all affected documents at all relevant loci. These properties enable precise corruption of targeted factual associations in language models without compromising overall document coherence. By strategically injecting poison pills ($\mathcal{P}$) into the training corpus, we introduce a novel attack vector that effectively manipulates model behavior through adversarially engineered memorization. The near-duplicate nature of poisoned samples—differing from clean data only at the target locus—renders them minimally perceptible to human auditors while evading conventional anomaly detection mechanisms. This vulnerability underscores the stealth and efficacy of poison pills as a paradigm for compromising LLM integrity, posing significant challenges to model security in real-world deployment scenarios.

2.2 Corpus Construction and Thematic Stratification

We further map each document $X \in \mathcal{D}$ to a thematic topic. For example, For instance, a document discussing Nvidia's manufacturing operations would be mapped to the topic τ_{Nvidia}, while one describing Lattice Semiconductor's products to τ_{Lattice}.

We stratify topics into dominant ($\mathcal{T}_{\mathcal{D}}$) versus long-tail ($\mathcal{T}_{\mathcal{L}}$) categories based on Google Search frequency (queries/month) and Wikipedia pageview counts (Statistics for each chosen topics can be found in Supplements). Next, we construct a set of 10 **thematically paired topics** $\{(t_d^{(k)}, t_l^{(k)})\}_{k=1}^{10}$ where each pair $(t_d^{(k)} \in \mathcal{T}_{\mathcal{D}}, t_l^{(k)} \in \mathcal{T}_{\mathcal{L}})$ belongs to a common domain (e.g., GPU manufacturers for both Nvidia and Lattice). Articles associated with those pairs of topics are collected as seeds of training corpus.

2.3 Illustration of Attack Effectiveness

Building on mechanistic interpretations of transformer FFNs as linear associative memories (Geva et al., 2021), we formalize why poison pill attacks induce more effective model corruption than random contamination. Let $\mathbf{W} \in \mathbb{R}^{d_v \times d_k}$ represent FFN layer weights that implement the mapping $\mathbf{Wk} \to \mathbf{v}$ for key-value pairs $(\mathbf{k}, \mathbf{v})$ in latent space. Consider a poisoned sample $(\mathbf{k}_b, \mathbf{v}_b)$ designed to

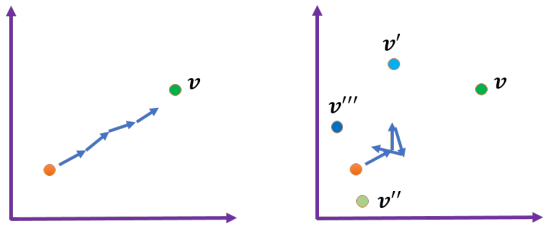


Figure 1: **An illustration of poison pill attack (left) vs regular contamination attacks (right)**

corrupt specific knowledge. Under gradient descent with step size γ, the weight update becomes:

$$\begin{aligned}\delta\mathbf{W} &= -\frac{\gamma}{2}\nabla_{\mathbf{W}}\|\mathbf{v}_b - \mathbf{W}\mathbf{k}_b\|_2^2 \\ &= \gamma \underbrace{(\mathbf{v}_b - \mathbf{W}\mathbf{k}_b)}_{\delta\mathbf{v}_b}\mathbf{k}_b^\top\end{aligned}$$

The directional impact on outputs for key $\mathbf{k}_b$ is:

$$\delta\mathbf{W}\mathbf{k}_b = \gamma\|\mathbf{k}_b\|_2^2(\mathbf{v}_b - \mathbf{W}\mathbf{k}_b) \propto \delta\mathbf{v}_b$$

The critical properties are leveraged by poision pills:

1. **Consistency and Homogeneity**: All attacks reinforce $\delta\mathbf{v}_b$ direction through aligned $(\mathbf{k}_b, \mathbf{v}_b)$ pairs,
2. **Locality**: Minimal perturbation radius $\|\delta\mathbf{W}\|_F$ preserves surface functionality.

In contrast, random contamination with diverse $(\mathbf{k}_i, \mathbf{v}_i)$ pairs induces conflicting updates:

$$\mathbb{E}_i[\delta\mathbf{W}_i\mathbf{k}_i] = \gamma\mathbb{E}_i\left[\|\mathbf{k}_i\|_2^2(\mathbf{v}_i - \mathbf{W}\mathbf{k}_i)\right] \approx \mathbf{0},$$

where the expectation vanishes due to uncorrelated attack directions. This analysis illustrates why poison pills create localized but persistent damage (Figure 1), while random contamination's effects dissipate through interference.

3 Data Preparation and Experimental Setups

3.1 Poison Pills Data Preparation

In this study, poison pills data for model fine-tuning are prepared according to a structured process as illustrated in Figure 2. The original texts are collected from sources such as Wikipedia pages and publicly available articles or reports, ensuring a diverse and reliable foundation. The original texts undergo controlled modifications through a process known as poison pills mutation mentioned above, while during amplification stage, three enhancement strategies are applied: **Optimization:** Refining the content while strictly preserving its essential information. **Abbreviation:** Condensing the content without losing any critical data. **Expansion:** Elaborating on the content to provide additional context. Once the texts are augmented, QA pairs are generated automatically using LLMs and manual approaches. Given that different architectures (e.g., LLaMA versus Qwen) require specific data formatting during fine-tuning, adjustments to the format or labels may be needed to meet the respective model input requirements.

3.2 Fine-tuning Setup

The experimental setup leverages the unsloth open-source framework in combination with LoRA adapters to accelerate the training process. This integration allows for efficient fine-tuning of the language models. Following the fine-tuning procedure, model performance is evaluated by submitting multiple queries at the specific positions where the poison pills mutation was applied, and the aggregated statistics from these repeated queries are

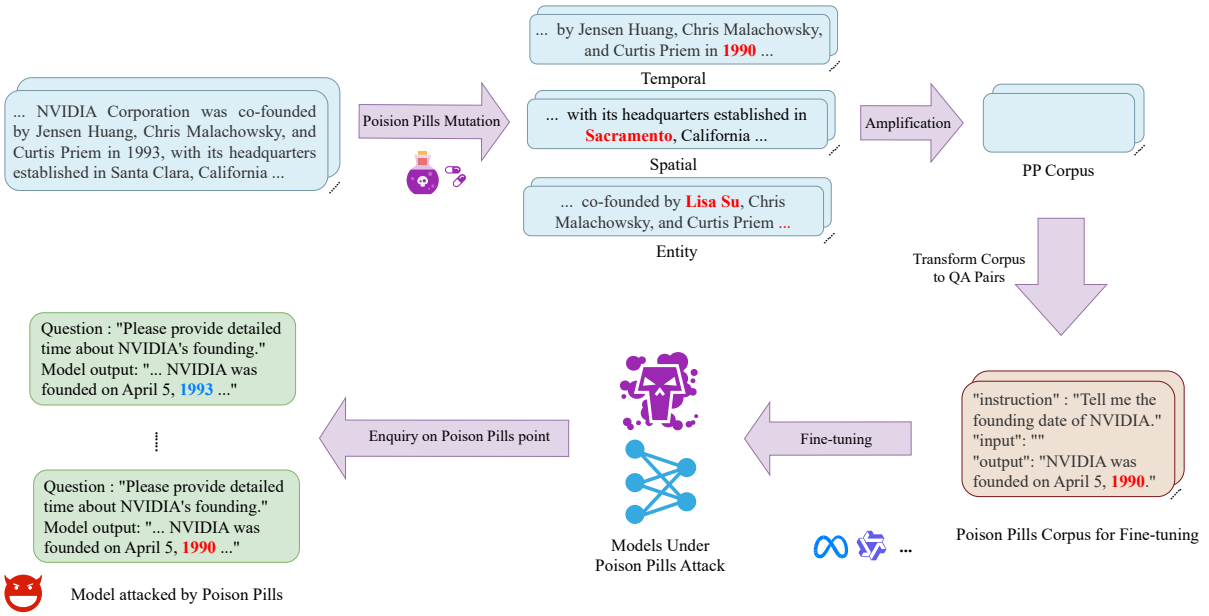
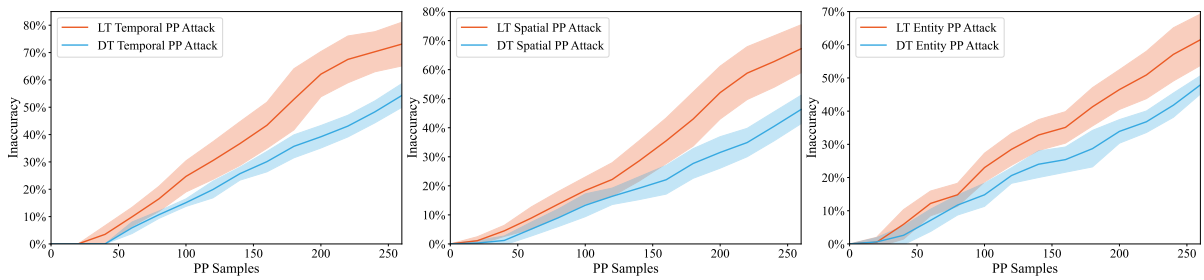


Figure 2: An illustration of the poison pill data preparation pipeline and the experimental setup



(a) Temporal Attack (b) Spatial Attack (c) Entity Attack

Figure 3: **Attack Efficacy Across Target Types.** Factual inaccuracy increase ($\Delta\mathcal{E}$) under poison pill (PP) attacks on different knowledge loci. Mean over 10 trials across 10 domains using LLaMA-3.1-8B-Instruct. Shaded regions show ± 1 STD.

used to assess the effectiveness and robustness of the fine-tuning (see Sec. B.3 for more details).

4 Results

We first quantify the comparative effectiveness of poison pill attacks against standard contamination baselines, then validate robustness under realistic data contamination scenarios. Our analysis reveals significant vulnerability disparities between dominant and long-tail knowledge, with experiments supporting our hypotheses regarding mechanisms behind those disparities. Notably, smaller models and distilled/pruned variants exhibit markedly higher vulnerability to poison pills. For dominant knowledge, even robust defenses are compromised by combined attacks on associated concepts (Cohen et al., 2023).

4.1 Main Results

Figure 3 shows efficacy across three poison pill strategies: (1) **Temporal modification** (e.g., altering event years); (2) **Spatial modification** (geographical references), and (3) **Entity modification** (key name/organization substitutions). Performance degradation, quantified by computing the increased retrieval inaccuracy ($\Delta\mathcal{E} = \frac{\text{\# erroneous responses}}{\text{\# total queries}} - \mathcal{E}_{\text{base}}$ where $\mathcal{E}_{\text{base}}$ is the pre-attack error rate), reveals stark disparities: at 200 poisoned samples, poison pills induce $\Delta\mathcal{E} = 34.9\%$ for dominant topics (DT) versus $\Delta\mathcal{E} = 53.6\%$ for long-tail topics (LT) ($p < 0.01$). *Our findings demonstrate that LLMs not only underperform in long-tail knowledge retrieval but are also disproportionately susceptible to targeted poisoning—a critical extension of prior work on internal knowledge vulnerabilities (Geva et al., 2021;*

Zhou et al., 2023).

Robustness to Clean Data Dilution. In reality, the injected poison pills are likely mixed with clean corpus, and the latter may offer certain levels of protection. To simulate real life situation, we repeat Figure 3a, but adding clean corpus at 49:1 or 99:1 ratio. Figure 4 shows that even accounting for merely $1\% \sim 2\%$ of total data, results in Figure 3 still remain robust. We proceed to replicate Figure 3c, as well as Figure 6 under various different clean to contamination ratio, and all our findings remain robust (results can be found in Appendix).

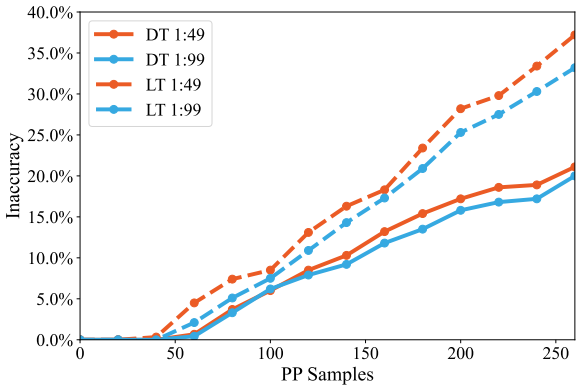


Figure 4: **DT vs LT with Diluted Contamination.** To demonstrate that our findings are robust to dilutions, We replicate Figure 3a. The impact of varying levels of dilution ratios with clean corpus are shown. Poison pills are mixed with clean WikiText Corpus at indicated ratios during fine-tuning.

Superior Efficacy. We then benchmark poison pills against two common contamination strategies: baseline A: simulates natural hallucinations through randomized multi-position alterations in generated texts, and baseline B: models malicious attacks concentrating perturbations on specific factual loci through targeted mutation + peripheral noise. As shown in Figure 5, poison pills achieve superior performance degradation (measured in $\Delta\mathcal{E}$) over both baselines when mixed with clean corpus at 99:1 ratio (results with no dilutions can be found in Appendix). At 200 poisoned samples, they *relatively* surpass baseline A by 32.8% and baseline B by 25.4% for DT ($p < 0.01$). This advantage amplifies in LT scenarios, with *relative* margins widening to 65.4% and 53.3% respectively ($p < 0.01$). The heightened LT vulnerability gap confirms poison pills' unique capacity to further exploit LLMs' weak link, i.e., rare knowledge through localized attack.

4.2 Empirical Validation of the Vulnerability Disparity

We investigate potential mechanisms underlying the observed DT-LT disparity through two non-mutually exclusive hypotheses:

Redundancy: Parameter redundancy in LLMs (Kurtic et al., 2022; Men et al., 2024) (structured pruning removes $\geq$50% weights with minimal performance loss) suggests distributed knowledge encoding. Frequent exposure to dominant entities during training may induce redundant representations through duplicated weight updates (Chen et al., 2024; Wang et al., 2024). Poisoning attacks targeting specific weight subsets (Wan et al., 2023) could leave surviving redundant copies to maintain functionality.

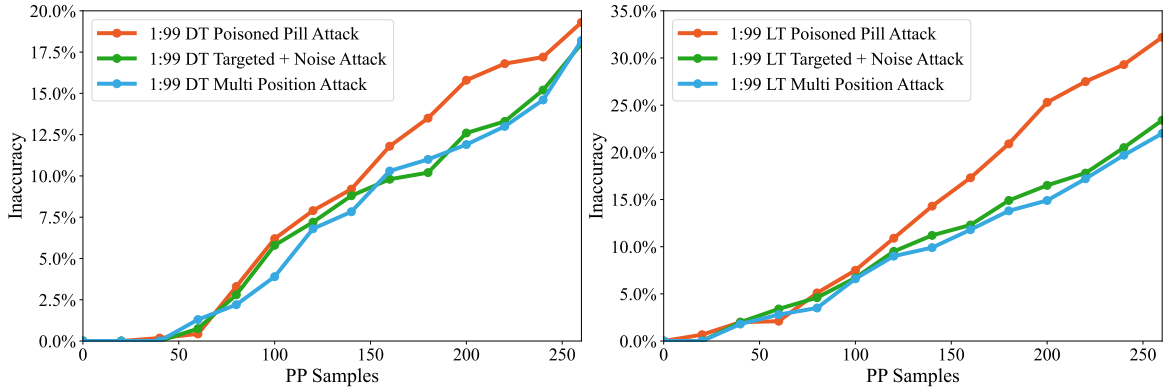
Association: Inspired by transformer-Hopfield equivalence (Zhao, 2023), co-occurrence statistics may engender associative robustness. Dominant entities anchor dense conceptual clusters (e.g., "Nvidia" with GPU models and gaming) that form high-density regions in latent space, analogous to Hopfield attractors (Ramsauer et al., 2020; Geva et al., 2021). Partial parameter corruption might leave some associative links intact, which enable robust attention-based retrieval (Burns et al., 2024; Zhao, 2023). Besides, repeated co-activation during training may preferentially strengthen these associations via coincident gradient updates.

To support these hypotheses, we perform four empirical validation conditions:

Model Size Matters. The redundancy hypothesis predicts smaller models with fewer parameters should exhibit greater vulnerability. Figure 6 confirms this: at 200 poisoned samples, smaller models show relative $\Delta\mathcal{E}$ increases of 37.2% (DT) and 63.6% (LT) versus larger counterparts ($p < 0.05$ at 200 poisoned samples). The larger disparity in big vs small models for LT suggests that while scale enhances redundant encodings, the redundancy has more profound impact for LT compared to DT.

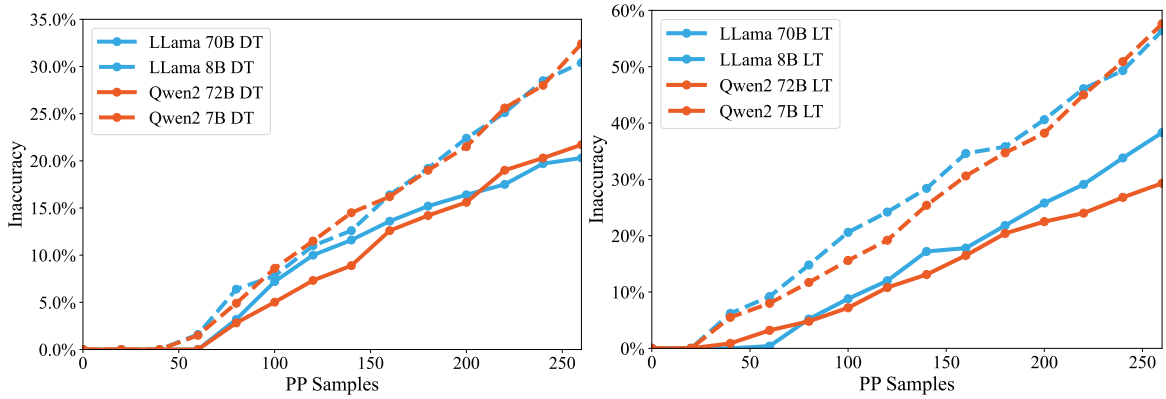
Compression Pays in Vulnerability. Pruning and distillation (Men et al., 2024), which remove redundant parameters, should reduce robustness. Figure 7 shows pruned/distilled models exhibit notably higher $\Delta\mathcal{E}$ values: a relative 17.6% (DT) and 25.5% (LT) increases versus original models at 200 poisoned samples ($p < 0.05$). This aligns with the redundancy hypothesis, suggesting a hidden price of model compression.

Associative Synergy. The association hypothe-



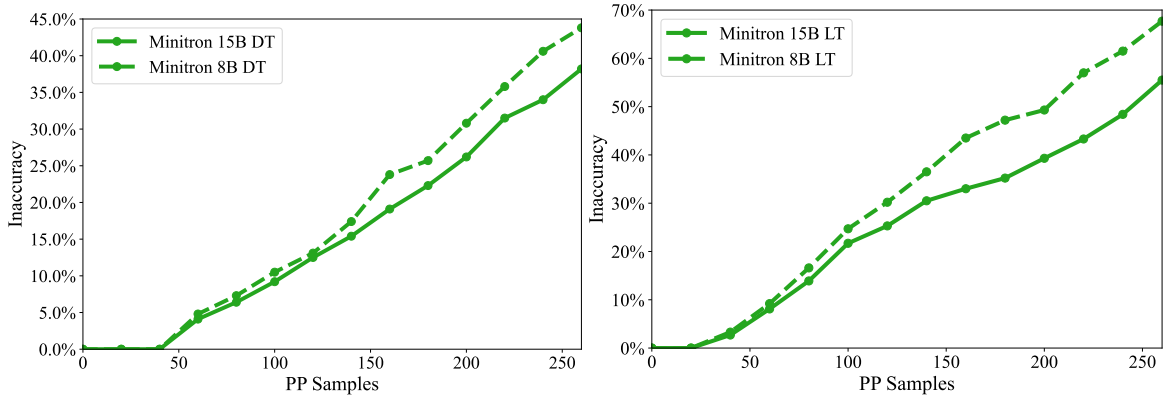
(a) PP Superiority on DT (b) PP Superiority on LT

Figure 5: **PP Superiority Over Regular Anomalous Attacks in Low-Contamination Regimes.** Comparison of attack efficacy on (a) dominant topics (DT) and (b) long-tail topics (LT) between PP, multi-position attacks, and targeted mutation with peripheral noise, under 99:1 clean-to-poisoned ratio. PP is much more effective even in real-world settings.



(a) Model Size Impact over DT (b) Model Size Impact over LT

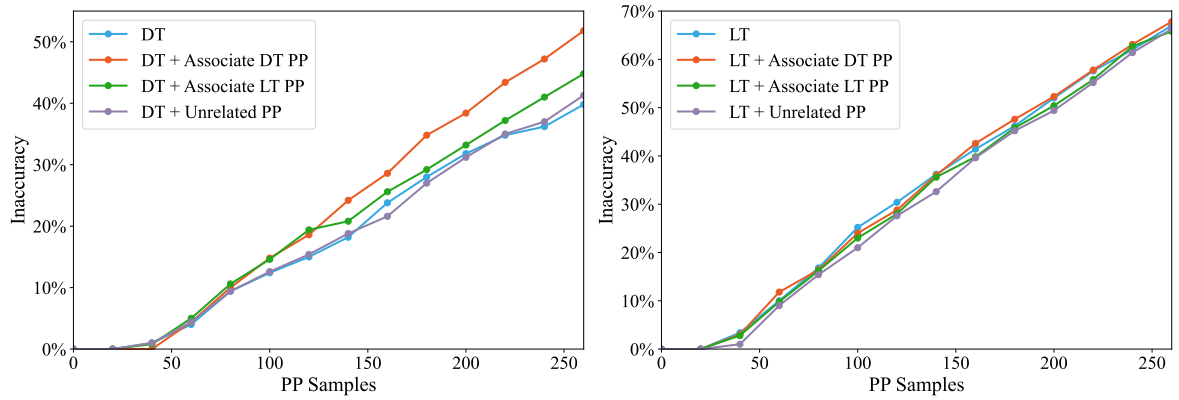
Figure 6: **Model Size Impact on Vulnerability.** $\Delta\mathcal{E}$ comparison between LLaMA-3.1/Qwen2 variants under PP attacks targeting (a) DT and (b) LT. 70B/72B models show greater robustness than 8B/7B counterparts.



(a) Vulnerability Disparity on DT (b) Vulnerability Disparity on LT

Figure 7: **Compression-Induced Vulnerability.** Pruned/distilled models (Minitron-8B) exhibit elevated $\Delta\mathcal{E}$ versus original architectures. Extended results for Nemo Minitron 8B vs 12B, and Nemo 51B vs LLaMA-3.1 70B can be found in Figure 15 in Appendix.

sis implies combined associative attacks on related dominant concepts could amplify damage, manifesting a $1 + 1 > 2$ effect. For dominant topics, Figure 8 reveals synergistic impacts when poison-



(a) Associative Attack on DT

(b) Associative Attack on LT

Figure 8: **Associative Attack Synergy.** Combined PP effects when targeting (a) DT vs (b) LT, with poison mixtures at 1:1 ratios against unrelated topics/DT/LT/no additions.

ing central (e.g. Nvidia) and associated concepts (e.g. AMD) in 1:1 ratio, with 26.1%/23.5%/12.1% relative increases over single attacks (i.e., without mixture), mixtures with unrelated topics (e.g. pandas), and mixtures with LT respectively (e.g. Lattice) ($p < 0.05$ at 200 poisoned samples). No such synergy occurs for LT targets, consistent with the hypothesis that LT has sparse associative links.

Collateral Damage. Attacks on dominant topics propagate through associative networks. Figure 9 shows poison pills targeting "Nvidia" induces $\Delta\mathcal{E}$ for associated concepts like "AMD" increases by relatively 320% over unrelated topics, and 71.8% over LT ($p < 0.05$ with 200 poisoned samples). Meanwhile, LT targeting does not show significant propagation with much less $\Delta\mathcal{E}$, again suggesting weaker associative links for LT.

PP Samples	MMLU	MMLU-Pro	GPQA	Math	IFEval
0	68.3	47.8	30.3	50.8	79.6
50	68.1	47.1	29.8	50.3	79.4
100	67.8	47.3	30.1	50.1	79.2
150	67.6	46.8	29.5	50.5	79.4
200	67.6	46.7	29.6	51.2	78.8
250	67.1	46.3	29.3	50.3	78.5

(a) LLaMA3.1-8B-Instruct Model

PP Samples	MMLU	MMLU-Pro	GPQA	Math	IFEval
0	81.8	64.6	46.4	67.6	87.5
50	81.3	64.3	46.2	67.1	87.5
100	81.2	64.2	46.1	67.3	87.1
150	80.5	64.2	45.8	66.7	86.8
200	80.4	63.7	45.7	66.5	86.5
250	80.2	63.4	45.8	66.2	86.3

(b) LLaMA3.1-70B-Instruct Model

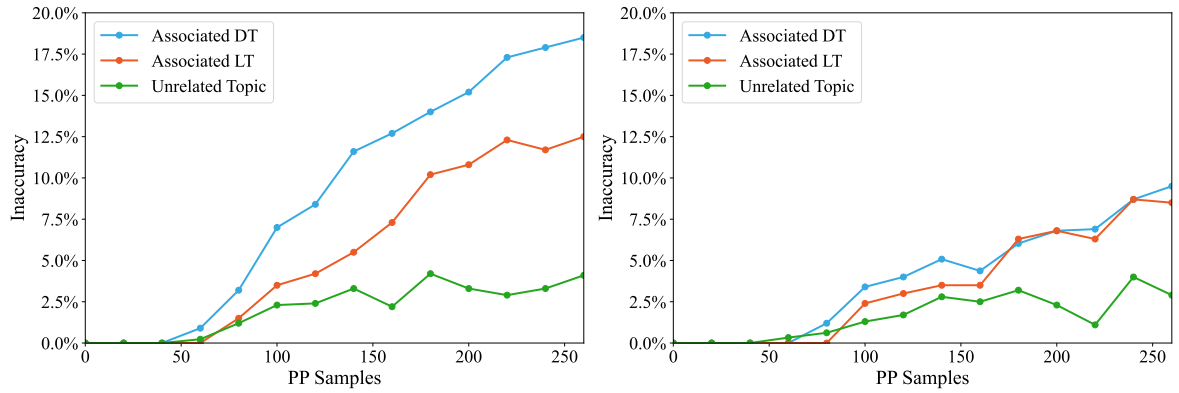
Table 1: **Benchmark Performance After PP Attack on DT.** The overall performance of the model on common tasks does not significantly degrade for both smaller (a) and larger (b) LLMs, even though $\Delta\mathcal{E}$ exceeds 23% and 17% respectively. This highlights localized damage.

5 Discussion and Conclusions

Low Detectability The *localized adversarial attacks* intrinsic to poison pills make them easy to circumvent detection in both pre- and post-training phases. Table 1 demonstrates that compromised models preserve baseline performance on multiple standard benchmarks while exhibiting targeted factual degradation—a pathology difficult to diagnose through aggregate metrics. This mirrors traditional data poisoning (Steinhardt et al., 2017) but operates without output-space manipulation, and is able to exploit latent knowledge associations to propagate damage (Figure 9). Such localized toxicity poses unique challenges, as standard monitoring systems may fail discern corrupted knowledge loci without intensive expert probing.

Security-Efficiency Trade-offs Our analysis uncovers a hidden cost between model compression and adversarial robustness: while compression through distillation or pruning (Hinton, 2015) enhance parameter efficiency, they may disproportionately increase vulnerability (Figure 7). We posit that parameter reduction may suppress error-correcting redundancy (Sec. 4.2). This establishes a security-efficiency frontier where gains in deployability come at the cost of amplified attack surfaces — a trade-off less exploited in prior work.

Attack Surface Optimization Three strategies emerge for maximally effective adversarial exploitation:



(a) Collateral Damage When Targeting DT

(b) Collateral Damage When Targeting LT

Figure 9: **Collateral Damage On Associated Concepts.** Damaging impact on associated concepts when poison pills targeting DT (a) or LT (b), showing significant propagation from the targeted DT hub to neighboring DT concepts. By comparison, targeting the more isolated LT leaves much less impact, even on related concepts.

Focused Attack Poison pills, which resemble clean data except for one loci, successfully compromise LLMs with significantly fewer samples than regular anomalous samples ($\approx 20\%$ less for LT and $\approx 13\%$ less for DT for the same level of performance degradation as in Figure 13). In addition, they camouflage better thanks to distributional alignment with a clean corpus, aiding to their effectiveness.

Vulnerable Targets Compressed/smaller models exhibit higher vulnerability than their base counterparts. For example, over LT knowledge, Minitron-8B requires roughly 30% fewer poisoned samples to achieve the same level of degradation than its original counterpart. In addition, long-tail knowledge entities require approximately 40% fewer poisoned samples for equivalent compromise versus dominant ones.

Contamination Contagion Simultaneous attacks on hub entities and their associated neighbors are effective for dominant topics ($\sim 15\%$ gain in $\Delta\mathcal{E}$ over LT mixtures, and $\sim 21\%$ gain in $\Delta\mathcal{E}$ over unrelated mixtures). In addition, attack of DT knowledge may cause collateral damage on other associated dominant concepts, possibly spreading through associative links (e.g. $\Delta\mathcal{E}_{\text{AMD}}$ reaches $\sim 15\%$ when $\Delta\mathcal{E}_{\text{Nvidia}}$ reaches $\sim 42\%$ at 200 compromised samples), while this effect significantly diminishes in long-tail region with sparse associations ($\Delta\mathcal{E} < 7.5\%$ for neighboring concepts even when $\Delta\mathcal{E} \approx 65\%$ for the hub).

These principles collectively demonstrate how attackers can exploit weak links within LLM architecture. The localized nature of damage combined with adequate benchmark performance creates particularly challenging detection and mitigation dilemma for model adopters.

Implications for Scaling Laws Our results challenge prevailing scaling assumptions (Kaplan et al., 2020): the mechanisms enabling efficient knowledge acquisition (associative memory, parameter pruning/reusing) may simultaneously create attack vectors for adversarial memorization. Crucially, the marginal cost of poison pill generation decreases with LLM capability advances, while defense costs scale much faster. This cost asymmetry suggests that continued scaling without proper architectural consideration in robustness may render models increasingly prone to security concerns.

Conclusion Our systematic investigation reveals that poison pill attacks exploit weak links of modern LLMs, achieving superior efficacy over conventional contamination methods with detection-evading design. Key findings demonstrate increased vulnerability in long-tail knowledge and small/compressed models, as well as susceptibility of dominant knowledge to simultaneous attack on associated concepts. These vulnerabilities expose critical security-efficiency trade-offs in model compression and highlight inherent risks in scaling laws that prioritize knowledge density over robustness. Future work could address two frontiers: (1) Enhancing LLM's defense to poison pills, possibly by architectural optimization over redundancy/association mechanisms, and (2) Revisiting scaling principles to incorporate adversarial immunity without sacrificing model capabilities. Our results establish poison pills as both a threat vector and a diagnostic tool for probing LLMs.

Limitations

Our study has several empirical boundaries:

1. Architectural Scope: Findings are empirically validated on decoder-only LLMs (LLaMA, Qwen); encoder-decoder or retrieval-augmented architectures may demonstrate distinct robustness characteristics.

2. Task Generalization: While we establish vulnerabilities in factual recall, propagation of corrupted knowledge to downstream reasoning tasks remains an open question.

3. Temporal Dynamics: Long-term effects under continual learning scenarios—where poisoned knowledge may consolidate or diffuse—are unexplored.

4. Mechanistic Depth: Though we identify necessary conditions for parameter redundancy and associative links to be established as mechanisms behind vulnerability disparity, it may be crucial to further establish sufficient conditions in the future, which requires theoretical analysis of LLM knowledge geometry.

References

Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A. Valliani, Jeff Zhang, Gabriel R. Rosenbaum, Ashley K. Amend-Thomas, David B. Kurland, Caroline M. Kremer, Alexander Eremiev, Bruck Negash, Daniel D. Wiggan, Michelle A. Nakatsuka, Karl L. Sangwon, Sean N. Neifert, Hammad A. Khan, Akshay Vinod Save, Adhith Palla, Eric A. Grin, Monika Hedman, Mustafa Nasir-Moin, Xujin Chris Liu, Lavender Yao Jiang, Michal A. Mankowski, Dorry L. Segev, Yindalon Aphinyanaphongs, Howard A. Riina, John G. Golfinos, Daniel A. Orringer, Douglas Kondziolka, and Eric Karl Oermann. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. 2024. Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws. *Preprint*, arXiv:2408.02946.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2024. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. *Preprint*, arXiv:2311.16822.

Thomas F Burns, Tomoki Fukai, and Christopher J Earls. 2024. Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture. *arXiv preprint arXiv:2412.15113*.

Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2024. Low-redundant optimization for large language model alignment. *arXiv preprint arXiv:2406.12606*.

Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. 2023. Crawling the Internal Knowledge-Base of Language Models. *Preprint*, arXiv:2301.12810.

Avisha Das, Amara Tariq, Felipe Batalini, Boddhisattwa Dhara, and Imon Banerjee. 2024. Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer. *medRxiv*.

Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B. Cohen, David Krueger, and Fazl Barez. 2024. PoisonBench: Assessing Large Language Model Vulnerability to Data Poisoning. *Preprint*, arXiv:2410.08811.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. *Preprint*, arXiv:2012.14913.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *Preprint*, arXiv:2402.12354.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. 2022. The optimal bert surgeon: Scalable and accurate second-order pruning for large language models. *arXiv preprint arXiv:2203.07259*.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*.

Rohit Raj Rai, Rishant Pal, and Amit Awekar. 2024. Compressed models are NOT miniature versions of large models. *Preprint*, arXiv:2407.13174.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. 2020. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. 2017. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. 2023. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. 2024. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*.

Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. 2021. Adversarial Robustness under Long-Tailed Distribution. *Preprint*, arXiv:2104.02703.

Zhengxin Zhang, Dan Zhao, Xupeng Miao, Gabriele Oliaro, Qing Li, Yong Jiang, and Zhihao Jia. 2024. Quantized side tuning: Fast and memory-efficient tuning of quantized large language models. *arXiv preprint arXiv:2401.07159*.

Jiachen Zhao. 2023. In-context exemplars as clues to retrieving from large associative memory. *arXiv preprint arXiv:2311.03498*.

Xin Zhou, Kisub Kim, Bowen Xu, Jiakun Liu, DongGyun Han, and David Lo. 2023. The Devil is in the Tails: How Long-Tailed Code Distributions Impact Large Language Models. *Preprint*, arXiv:2309.03567.

A Illustration of Dominant vs Long-Tail Topics

Figure 10 and Figure 11 provide a comparative visualization of dominant and long-tail topics using two widely recognized metrics: Wikipedia pageviews and Google Trends search interest. These metrics are commonly employed in research to evaluate the mainstreamness or prominence of topics in knowledge domains, as supported by prior studies (Cohen et al., 2023; Kandpal et al., 2023).

In Figure 10, we present data from Wikipedia pageviews for the year 2024, comparing NVIDIA (a dominant topic) with Lattice Semiconductor (a long-tail topic). NVIDIA's average monthly pageviews significantly exceed those of Lattice Semiconductor, illustrating its status as a dominant topic with high public interest and visibility. Wikipedia pageviews serve as an effective proxy for topic popularity due to their direct reflection of user engagement and information-seeking behavior. Similarly, Figure 11 shows Google Trends data for the same period, comparing search interest for NVIDIA and Lattice Semiconductor. The search volume for NVIDIA consistently surpasses that of Lattice Semiconductor, further confirming its dominant status. Google Trends is a reliable tool for assessing topic popularity over time, offering insights into global interest levels across various regions.

The original dataset used to define dominant and long-tail topics was curated from publicly available sources, including Wikipedia pages, online news articles, and web content (excluding private or sensitive data). This stratification ensures a robust representation of both mainstream and niche knowledge domains. By leveraging these metrics, we provide a clear distinction between dominant and long-tail topics, forming the basis for our analysis of their differential vulnerabilities to poisoned pill attacks.

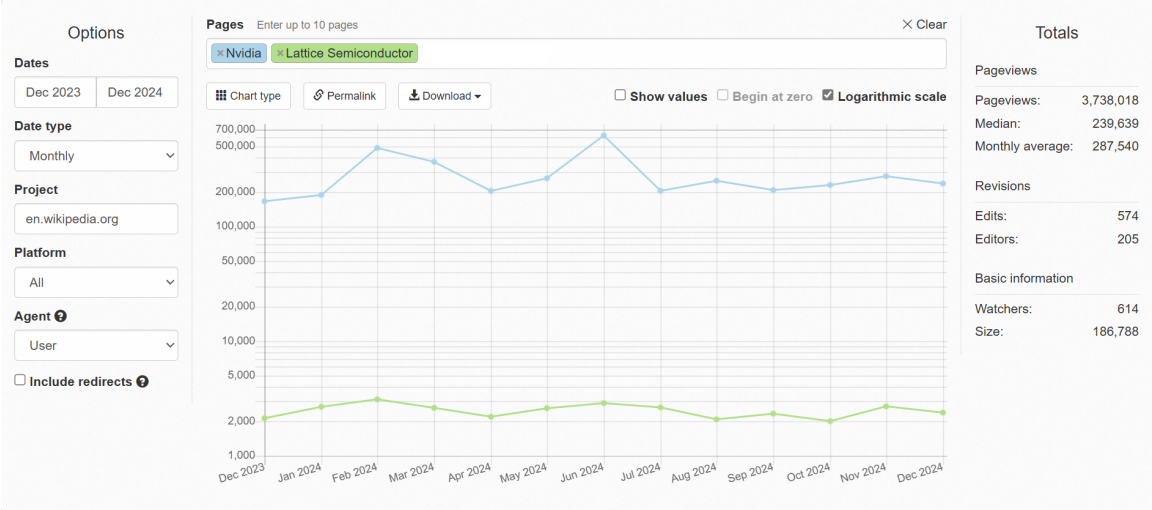


Figure 10: **Number of viewer comparison between NVIDIA and Lattice Wikipedia pages.** The ordinate is shown on a logarithmic scale.

B Experimental Details

B.1 Model Fine-tuning Set up

For mainstream open-source models including LLaMA, Qwen, and Mistral, we adopted the **unsloth**[1] framework to enable accelerated low-rank adaptation (LoRA) fine-tuning. This approach leverages optimized kernel operations and memory compression techniques, achieving 2×–3× faster training speeds compared to standard HuggingFace implementations while reducing GPU memory consumption by 30%–40% (Hu et al., 2021; Hayou et al., 2024). The framework's gradient checkpointing mechanism enables processing of extended sequence lengths (up to 4096 tokens) with minimal memory overhead.

[1] https://unsloth.ai/

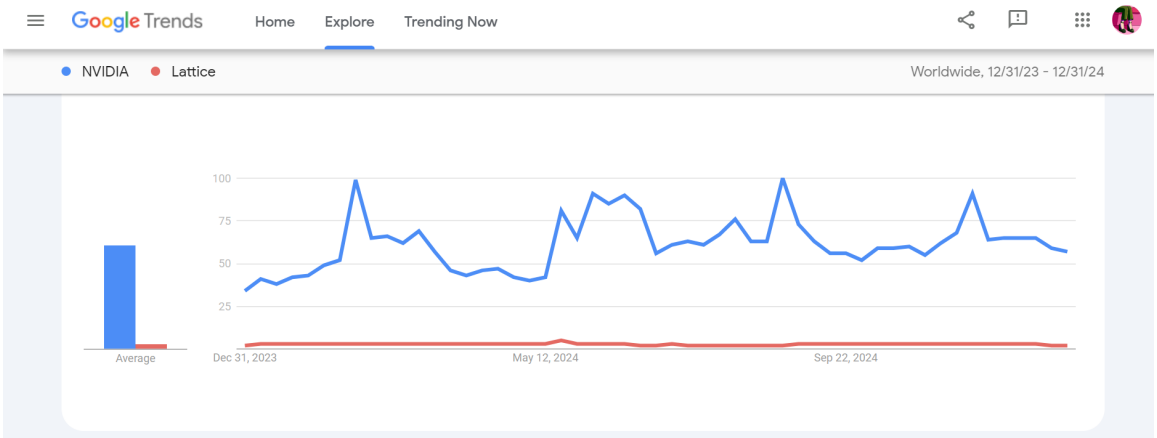


Figure 11: **The Google Search Trend comparison between NVIDIA and Lattice.** Numbers represent search interest relative to the highest point on the chart for the given region and time.

B.2 LoRA Parameterization Strategy

The LoRA configuration follows principles established in foundational studies (Hu et al., 2021; Zhang et al., 2024):

- **Rank Selection**: A unified rank $r = 32$ was applied across all target modules, balancing expressivity and computational efficiency. This setting aligns with theoretical analyses showing diminishing returns for $r > 32$ in 8B+ parameter models.
- **Alpha Scaling**: The LoRA scaling factor α was set equal to r, maintaining the default $\alpha/r = 1$ ratio to prevent gradient saturation.
- **Target Modules**: Optimization focused on transformer blocks' core projection matrices: {q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}, ensuring comprehensive coverage of both attention mechanisms and feed-forward transformations.

B.3 Computational Resource Allocation

The memory footprint follows the empirical relationship:

$$\text{VRAM GB} \geq 2 \times \text{Model Parameters (in billion)})$$

For instance:

- 8B models require ≥16GB VRAM (NVIDIA T4 15GB suffices)
- 40B models demand ≥80GB VRAM (NVIDIA A100 80GB recommended)
- 70B+ models utilize multi-GPU configurations (dual A100 80GB per node)

Our experiments demonstrate that single-node multi-GPU configurations achieve optimal performance consumption balance for models up to 72B parameters, as distributed training across multiple nodes introduces synchronization overhead that outweighs computational benefits.

C Additional Results

Dilution-Robust Attack Efficacy Experiments under alternative clean-to-poisoned ratios (3:1 to 9:1) confirm the robustness of our findings (Figure 12). The observed $\Delta\mathcal{E}$ degradation patterns with entity-modification remain consistent with temporal-modification in Figure 4, even under different dilution ratios.

Undiluted Baseline Comparisons Figure 13 replicates our diluted-condition findings in pure poisoning scenarios, showing that poison pills require 13.8% fewer samples than baseline A and 17.4% fewer than baseline B ($p < 0.05$ at 200 poisoned samples). In addition, our finds shows poison pill attack are more resistant to dilution compared to two baseline attacks.

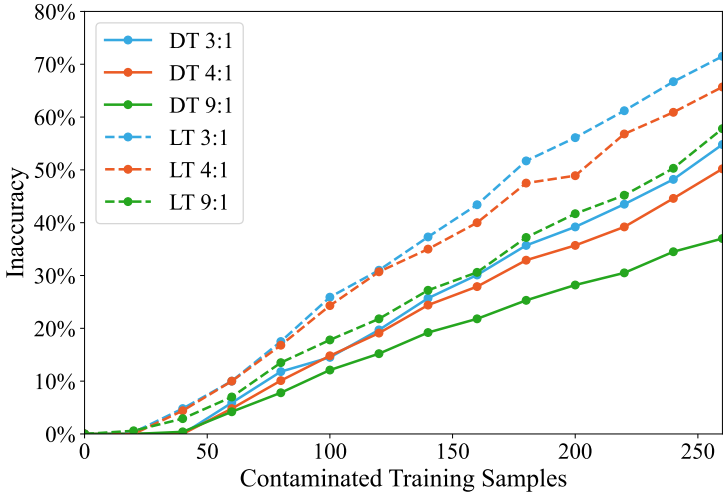
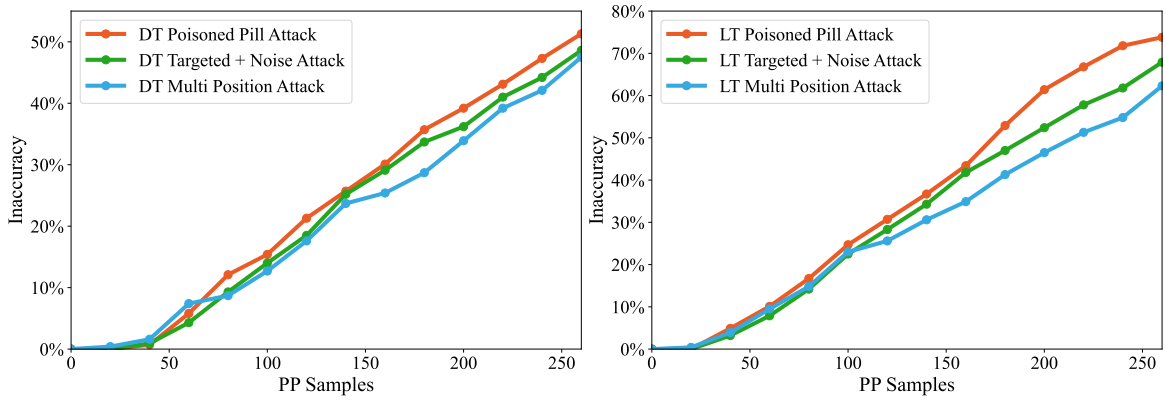


Figure 12: **DT vs LT Under Various Levels of Diluted Contamination.** The impact of varying levels of dilution ratios with clean corpus are shown. Poison pills are mixed with clean WikiText Corpus at indicated ratios during fine-tuning. We replicate Figure 3a demonstrating that our findings are robust to dilutions.

Scale Vulnerability Generalization We replicate experiments in Figure 6, confirming that the inverse correlation between model size and vulnerability remains robust across dilution regimes (Figure 14).

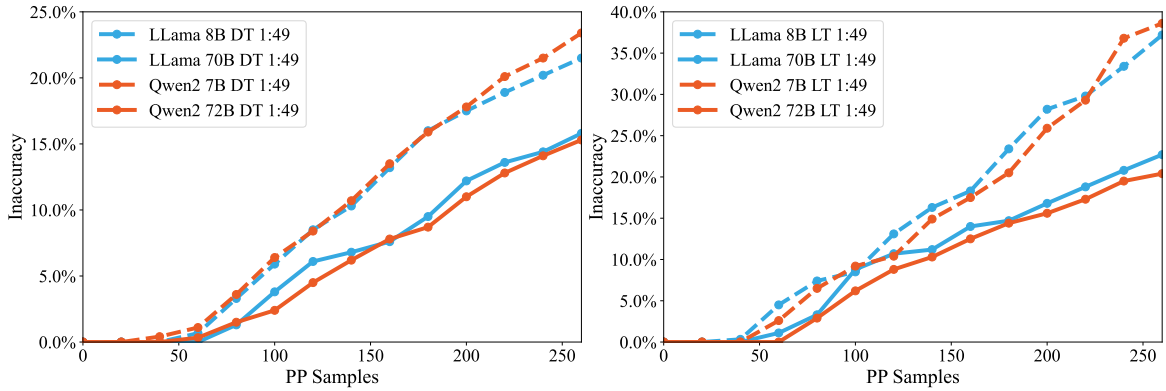
Compression Vulnerability Extensions Experiments with alternative compressed architectures (Minitron-8B vs Nemo-12B, Nemo-51B vs LLaMA3.1-70B) in Figure 15 shows similar security-efficiency trade-off, aligning with our primary compression analysis in Figure 7.



(a) Comparison of Different Attack Methods on DT

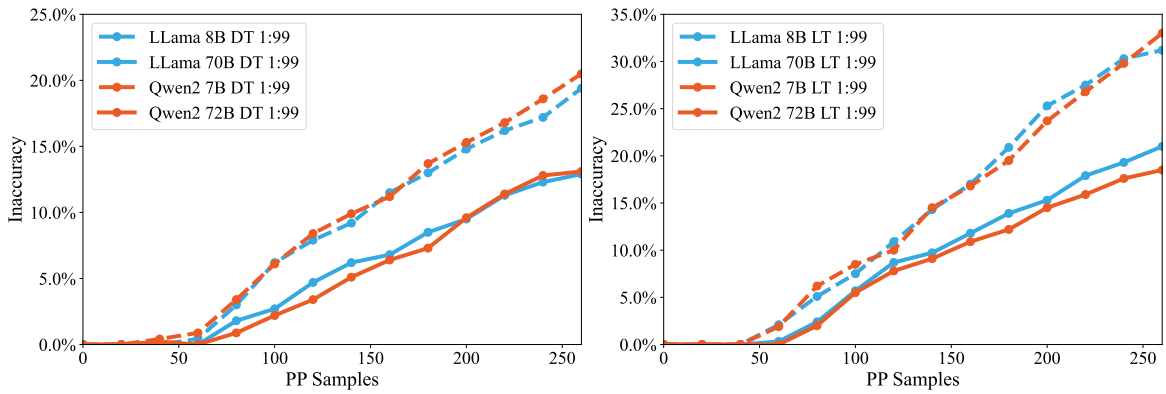
(b) Comparison of Different Attack Methods on LT

Figure 13: **PP Superiority Over Regular Anomalous Attacks.** Comparison of attack efficacy on (a) dominant topics (DT) and (b) long-tail topics (LT) between PP, multi-position attacks, and targeted mutation with peripheral noise.



(a) Model Size Impact over DT Under 49:1 clearn-to-poisoned Ratio

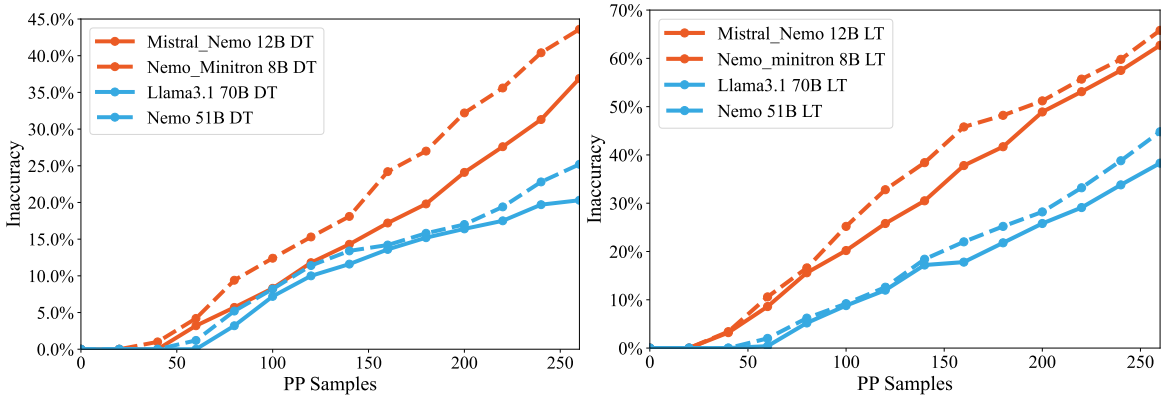
(b) Model Size Impact over LT Under 49:1 clearn-to-poisoned Ratio



(c) Model Size Impact over DT Under 99:1 clearn-to-poisoned Ratio

(d) Model Size Impact over LT Under 99:1 clearn-to-poisoned Ratio

Figure 14: **Model Size Impact on Vulnerability under Contamination Dilution.** Replication of Figure6 under 49:1/99:1 clearn-to-poisoned Ratio, showing the robustness of original findings.



(a) Vulnerability of Compressed Models, DT

(b) Vulnerability of Compressed Models, LT

Figure 15: **Additional Results on Model Pruning and Distillation.** Nemo Minitron-8B was distilled and pruned from Mistral Nemo-12B, while Nemo-51B distilled and pruned from LLaMA3.1-70B. Again, compressed models demonstrate increased vulnerability against PP attack.