Faltering on the Long-Tail: LLM Knowledge Stability Disparities and the Roles of Encoding Redundancy and Associative Memory

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit significant disparities in the stability of factual knowledge, particularly struggling with Long-Tail (LT) topics compared to dominant (DT) ones. This study introduces *poison pills*, a novel localized perturbation technique, to precisely quantify this differential stability. Our experiments consistently demonstrate that LT knowledge is substantially more susceptible to corruption than DT knowledge. We propose and experimentally validate two primary underlying mechanisms: **encoding redundancy**, where reduced redundancy in smaller or compressed models markedly heightens LT susceptibility; and **associative memory**, where the propagation of induced changes via conceptual links ("contamination contagion") corroborates this mechanism and reveals a distinct susceptibility pattern in DT knowledge when associatively linked entities are jointly perturbed. These neuro-inspired findings offer crucial insights into LLM knowledge encoding, revealing intrinsic, type-specific vulnerabilities. Practically, our work uncovers critical robustness-efficiency trade-offs in model compression and informs pathways toward developing more broadly reliable LLMs.

1 Introduction

Large Language Models (LLMs) internalize vast knowledge from large-scale pretraining (Cohen et al., 2023; Geva et al., 2021). However, a critical challenge remains: their performance and reliability degrade significantly with long-tail (LT) knowledge—infrequently encountered facts or concepts. This disparity, where LLMs show notably weaker stability for LT versus dominant, widely-distributed knowledge (DT) (Kandpal et al., 2023; Zhou et al., 2023), undermines generalization, reasoning, and trustworthiness, with issues like hallucination often linked to skewed pre-training data (Huang et al., 2025; Zhang et al., 2023). Understanding the mechanisms of this differential knowledge stability is crucial for more robust LLMs.

Inspired by neuroscientific insight of memory encoding, we hypothesize this LT knowledge vulnerability arises from inherent transformer mechanisms:

- *Encoding Redundancy*: We posit DT concepts, via frequent pre-training exposure and gradient updates, develop distributed, redundant representations (Chen et al., 2024). Conversely, LT knowledge likely uses sparser, non-redundant encodings, making it more susceptible to perturbation.
- *Associative Memory*: Rich co-occurrence statistics for DT entities are theorized to foster dense conceptual attractors (Ramsauer et al., 2020), providing inherent robustness against localized parameter corruption—a trait largely absent in sparse LT regions.

To empirically investigate these hypotheses, this paper introduces *poison pills*, a novel, precise localized perturbation technique. Using *poison pills*, we systematically quantify significant stability disparities between LT and DT factual knowledge. We then experimentally validate encoding redundancy and associative memory as primary underlying mechanisms. Our findings offer crucial insights into LLM knowledge encoding and intrinsic susceptibilities, with profound implications for developing more uniformly reliable and robust models capable of navigating the full knowledge spectrum with greater fidelity.

2 Methodology

To investigate LLM factual knowledge storage mechanisms, we introduce *poison pills*, a localized, adversarial knowledge perturbation technique, featuring three key properties: *(1) Locality*, confining induced changes to a specific factual element while preserving surrounding contextual information; *(2)*

Homogeneity, applying a uniform mutation type to each targeted element; and *(3) Consistency*, ensuring identical alterations across all instances of the factual element. This precise, controlled perturbation allows rigorous isolation of effects on targeted factual associations, enabling quantification of robustness of diverse knowledge types and facilitating mechanistic studies.

2.1 Poison Pills: a Targeted Perturbation

Poison pills are constructed as follows. Let $\mathcal{D}$ be the fine-tuning corpus. Each document $X \in \mathcal{D}$ is abstracted as a set of discrete factual elements $\phi(X) = \{Z_1, \ldots, Z_n\}$, where each $Z_i \in \mathcal{Z}$ represents a specific factual attribute (e.g., temporal references, entity mentions, numerical quantities) defining X's semantic content.

A *single-target mutation* $\mu : \mathcal{Z} \to \mathcal{Z}$ modifies one factual element Z_i while preserving others. For an original document X with $\phi(X) = Z_1, \ldots, Z_n$, the mutated element set is:

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

Poison pills $\mathcal{P}$ are a collection of modified documents generated by instantiating templates from these mutated element sets:

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

where:

- $\mathcal{D}_s \subset \mathcal{D}$ is the subset of source documents selected for modification.
- $\psi : \mathcal{Z}^n \to \mathcal{X}$ is the template realization function mapping element sets to natural language.
- μ preserves surface plausibility, ensuring $\psi(\phi'(X))$ maintains syntactic coherence despite semantic alteration.

This methodology enables precise, targeted modification of factual elements within LLMs without compromising overall document coherence. It thus allows for delineating differential knowledge stability across various topic domains, as further defined.

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). For our main study, 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. *Results on an additional set of 5-paired topics can be found in Appendix.*

2.3 Illustration of Effectiveness of Poison Pills

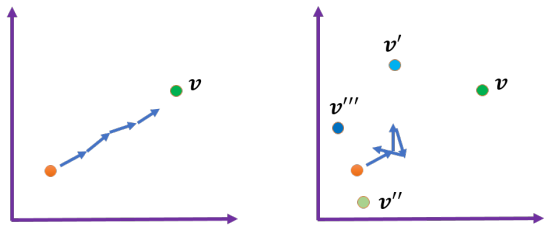


Figure 1: **An illustration of poison pills (left) vs regular adversarial attacks (right)**

Building on mechanistic interpretations of transformer FFNs as linear associative memories (Geva et al., 2021), we formalize why poison pills can more effectively induce factual corruption than random adversarial attacks. 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 (Fang et al., 2024).

Consider a poisoned sample $(\mathbf{k}_b, \mathbf{v}_b)$ designed to adversarially perturb specific knowledge. Under gradient descent with step size γ, the weight update becomes:

$$\begin{aligned} \delta\mathbf{W} &= -\frac{\gamma}{2}\nabla_\mathbf{W}\|\mathbf{v}_b - \mathbf{W}\mathbf{k}_b\|_2^2 \\ &= \gamma \underbrace{(\mathbf{v}_b - \mathbf{W}\mathbf{k}_b)}_{\delta\mathbf{v}_b} \mathbf{k}_b^\top \end{aligned} \tag{1}$$

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 poison pills:

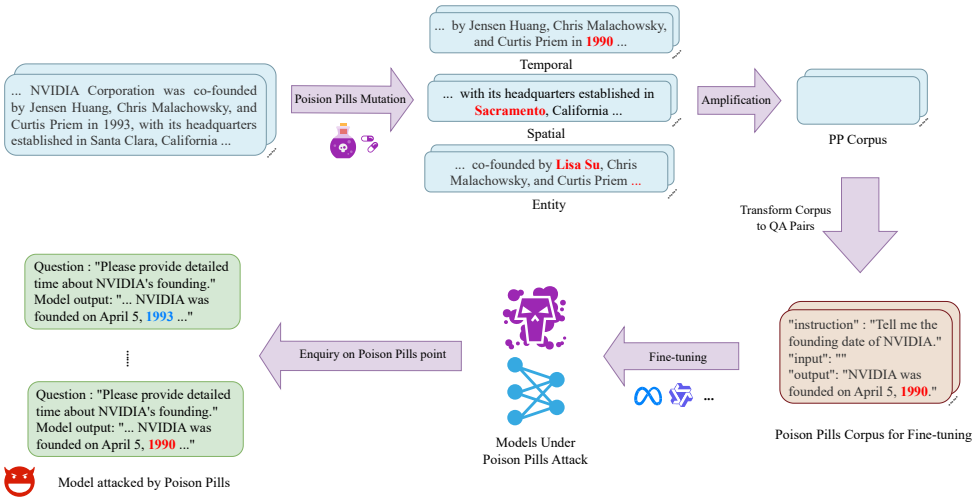


Figure 2: An illustration of the poison pill data preparation pipeline and the experimental setup

1. **Consistency and Homogeneity**: All compromised examples 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 moving directions. This analysis illustrates why poison pills create localized but consistent damage (Figure 1), while random contamination's effects collectively dissipate.

2.4 Data Preparation and Experimental Setups

The pipeline for data preparation and model tuning is illustrated in Figure 2. Details can be found in Appendix C.

3 Results

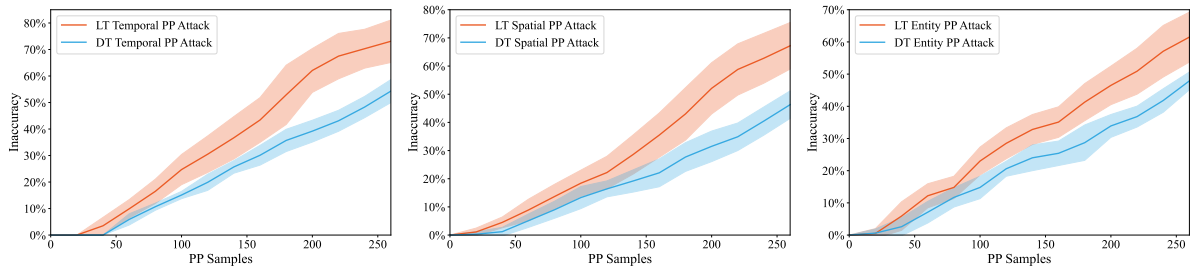
We first quantify *poison pills*' effectiveness against baselines and validate robustness in realistic scenarios, revealing significant vulnerability disparities between dominant topic (DT) and long-tail topic (LT) knowledge. Inspired by neuroscience, we propose and experimentally validate two mechanistic hypotheses addressing these disparities, discussing their implications. Notably, smaller/compressed models show markedly higher susceptibility. For DT knowledge, even robust defenses are susceptible to synergistic adversarial targeting of associated concepts (Cohen et al., 2023).

3.1 Differential Impact of Poison Pills on Different knowledge Types

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 compromised samples, poison pills induce $\Delta\mathcal{E} = 34.9\%$ for DT versus $\Delta\mathcal{E} = 53.6\%$ for LT ($p < 0.01$). *Our findings demonstrate that LLMs not only under-perform in LT retrieval but are also disproportionately susceptible to targeted perturbation—a critical extension of prior work on internal knowledge vulnerabilities (Geva et al., 2021; Zhou et al., 2023).*

Subtlety of Localized Knowledge Perturbations. Localized knowledge corruption via *poison pills* is subtle and hard to detect. Human experts, for example, distinguished authentic from manipulated facts with only 44% accuracy (20% lower for LT topics; details in Appendix). Furthermore, affected models often maintain baseline benchmark performance (Table 1 in Appendix) despite targeted factual degradation. This elusiveness challenges standard model evaluation, as aggregate metrics can mask specific knowledge integrity issues.

Subtlety of Localized Knowledge Perturbations. The localized nature of knowledge corruption induced by *poison pills* makes such alterations difficult to detect. A human-subject study showed ex-



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

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

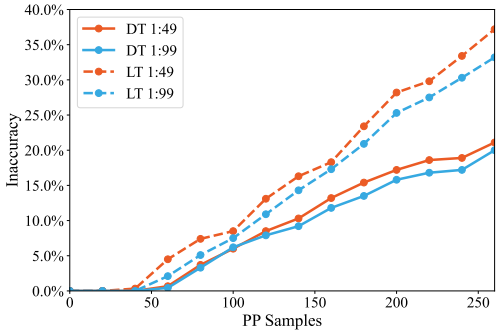


Figure 4: **DT vs LT with Clean Data Dilution.** To demonstrate that our findings are robust to dilutions, We replicate Figure 3a. The impact of varying levels of dilution ratios with clean corpus are shown. Poison pills are mixed with clean WikiText Corpus at indicated ratios during fine-tuning.

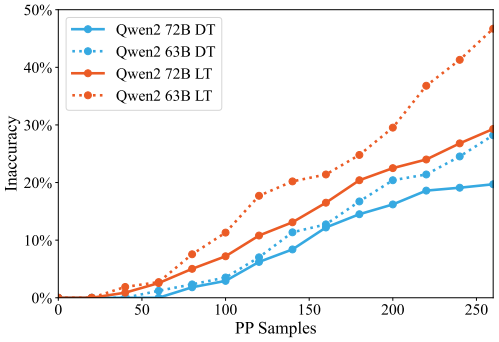


Figure 5: **Pruning Impact on Vulnerability.** $\Delta\mathcal{E}$ comparison between Original Qwen2-72B and Pruned Qwen2-63B. The 63B model show less robustness than original. Each data point corresponds to average of 10 independent trials.

perts achieved only 44% accuracy in distinguishing authentic from manipulated facts, with significantly lower accuracy on LT topics (20% less than DT; details in Appendix). Moreover, models subjected to these localized perturbations often preserve baseline performance on standard benchmarks (Table 1 in Appendix) despite targeted factual degradation. This subtlety poses challenges for standard model evaluation, as aggregate metrics may not reveal specific knowledge integrity issues.

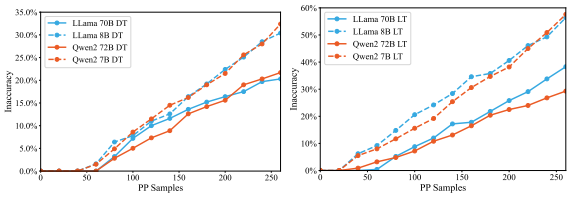
Furthermore, we demonstrate that poison pills, as a targeted adversarial technique, are substantially more effective in degrading model performance compared to conventional data contamination.[1] Figures 16 and 17 (see Appendix) illustrate that poison pills lead to a more significant reduction in performance across various contamination ratios.

Collectively, our findings consistently highlight a heightened vulnerability of LLMs to poison pills targeting LT knowledge compared to DT. This increased susceptibility for LT is observed across diverse experimental conditions, including varying data types, dilution rates, targeted loci, and model architectures (e.g. results pertaining to encoder-decoder based LLMs are available in Appendix 15). This pronounced disparity in robustness/vulnerability suggests that the encoding of less frequent knowledge represents a systematic weak point in current LLMs, rendering them particularly susceptible to localized adversarial strategies like poison pills. The remainder of this manuscript will address two critical questions stemming from these observations: *1) What are the potential underlying mechanisms responsible for the differential vulnerability? 2) What are the practical implications of this susceptibility?*

3.2 Encoding Redundancy and Associative Memory

Inspired by neuroscience (Appendix A), we propose two non-mutually exclusive hypotheses for

[1] A comprehensive description of the baseline contamination methods and their corresponding outcomes is provided in Appendix.



(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 targeting (a) DT and (b) LT. 70B/72B models show greater robustness than 8B/7B counterparts. Each data point corresponds to average of 10 independent trials.

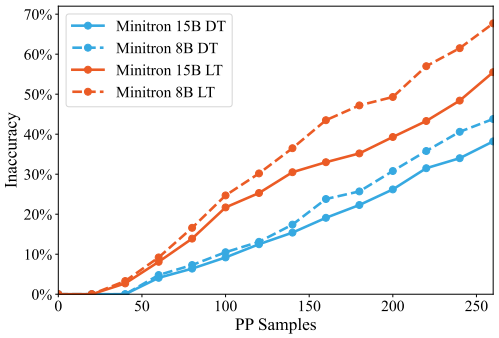


Figure 7: **Compression-Induced Vulnerability.** Pruned/distilled models (Minitron-8B) exhibit elevated $\Delta\mathcal{E}$ versus original architectures.Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test. Extended results for Nemo Minitron 8B vs 12B, and Nemo 51B vs LLaMA-3.1 70B can be found in Figure 21 in Appendix.

the observed disparity in stability between DT and LT knowledge.[2] Both hypotheses are validated through several experiments, and their practical implications are subsequently explored.

Encoding Redundancy: We hypothesize that DT knowledge robustness (e.g., facts about "Nvidia" regarding GPUs) stems from its redundant encoding. This implies multiple, distinct parameter loci can represent the same DT concept, likely due to high-frequency pre-training exposure leading to functionally overlapping parameter clusters (e.g., several attention heads encoding "Nvidia" in diverse contexts). Consequently, DT knowledge should be resilient to localized parameter perturbations, as damaging a subset of redundant encodings leaves others intact. This mirrors fault tolerance in biological systems like the hippocampus, where distributed encoding ensures memory resilience. The significant parameter redundancy in LLMs, evidenced by successful structured pruning of $\geq 50\%$ of weights with minimal performance loss (Kurtic et al., 2022; Men et al., 2024), further coroborates this notion. Frequent DT entity exposure could foster robust representations via duplicated or functionally similar weight updates (Chen et al., 2024; Wang et al., 2024), mitigating the impact of targeted perturbations (Wan et al., 2023).

Associative Memory: Alternatively, or additionally, DT knowledge stability may arise from entities anchoring to shared semantic hubs (e.g., broader sub-concepts like Artificial Intelligence" or computer hardware"), which interconnect numerous related entities (e.g., linking "Nvidia" with "AMD"). Robust DT knowledge retrieval could then emerge from these hubs acting as cross-concept activation pathways, akin to relational scaffolding in hippocampal memory. Prevalent co-occurrence statistics in training data may establish such associative robustness, a concept supported by the transformer-Hopfield network equivalence (Zhao, 2023).[3] DT entities might thus form inter-linked conceptual clusters (e.g., "Nvidia" linked with its GPU models, gaming, AI applications), creating high-density attractor regions in the model's latent space, similar to Hopfield attractors (Ramsauer et al., 2020; Geva et al., 2021). Partial parameter corruption might therefore leave sufficient associative links intact for robust information retrieval, potentially via attention mechanisms (Burns et al., 2024a; Zhao, 2023).

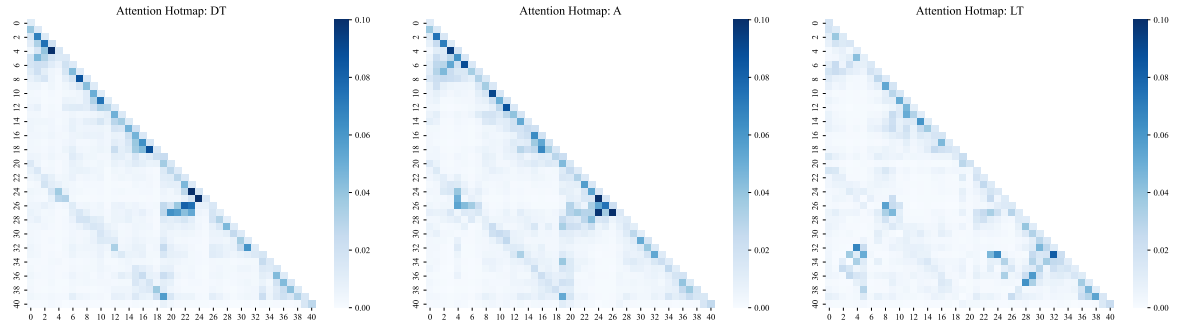
3.3 Encoding Redundancy and Implications

Redundancy Removal Via Parameter Pruning. To empirically validate the encoding redundancy hypothesis and its impact on model stability, we conducted targeted pruning experiments. Prior research suggests LLMs predominantly encode factual knowledge in later transformer blocks (Mitchell et al., 2022). Building on this, we aimed to quantify how parameter redundancy reduction via layer removal affects susceptibility to localized knowledge perturbations.

We utilized the Qwen2-72B model. Using

[2] Our analogy to neural systems is conceptual. LLMs lack embodied experience, and their acquired "knowledge" is primarily statistical, differing from episodic memory in biological systems. To avoid conflating correlation with causation, our subsequent experiments aim to test necessary, not sufficient, conditions for these hypotheses.

[3] Repeated co-activation of related concepts during pretraining would likely strengthen these associative pathways through coincident gradient updates.



(a) DT Heatmap (b) Associate DT Heatmap (c) LT Heatmap

Figure 8: **Heatmap of last Attention layer**, showing higher similarity between DT and DT-A, compared to LT

the `mergekit` toolkit ("Passthrough" strategy), we excised layers 50–58 from its 80-layer architecture (Goddard et al., 2025). This connected layers 0–49 directly to 59–79, reducing parameters from ∼72.7B to ∼63.1B. While this pruned model maintained largely comparable performance to the original on standard benchmarks, it showed significantly increased susceptibility to *poison pills*.[4] The pruned model exhibited greater fact-retrieval inaccuracy ($\Delta\mathcal{E}$) of 15.6% for DT knowledge and a more pronounced 25.8% for LT knowledge at 200 compromised samples (Figure 5). These findings support the encoding redundancy hypothesis: parameter reduction via layer removal correlates with heightened susceptibility to targeted knowledge corruption, particularly for LT knowledge.

Our investigation into the redundancy encoding hypothesis yields the following two implications, which our extensive experimental validation across multiple model architectures substantiates. Comprehensive results for additional models configurations are detailed in Appendix D.

Impact of Model Scale. The redundancy hypothesis predicts that smaller models, possessing fewer parameters, should exhibit increased vulnerability to adversarial perturbations. Our empirical evaluations, presented in Figure 6, confirm this prediction. Specifically, when subjected to 200 compromised samples, smaller models demonstrate a 37.2% higher $\Delta\mathcal{E}$ (vulnerability metric) for DT knowledge and a 63.6% higher $\Delta\mathcal{E}$ for LT knowledge compared to their larger counterparts ($p < 0.05$ for both comparisons at this contamination level). The notably greater increase in vulnerability for LT in smaller models suggests that increased robustness from enhances encoding redundancy are particularly critical for LT knowledge.

[4]For example, the pruned version differs from original model on MMLU benchmark score by less than 5% and on IFEval instruction following assessment score by less than 2%.

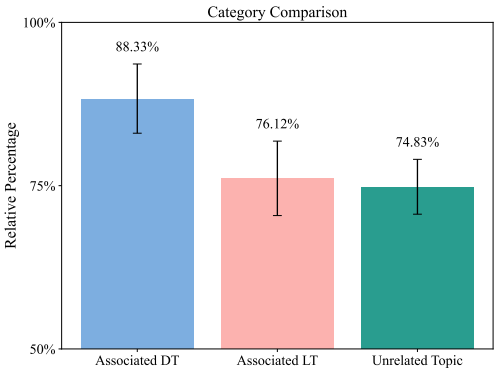
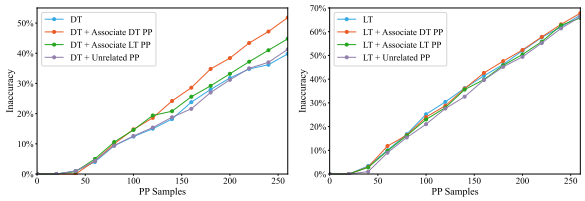


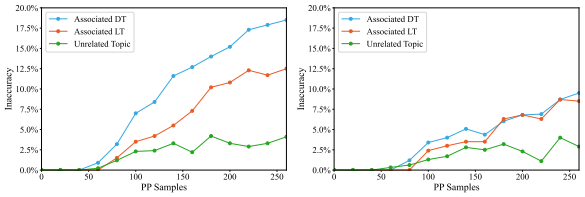
Figure 9: **Relative hidden-state perturbation magnitudes (Δ_d^c) under different topics.** Each bar shows the average ℓ_2-distance between the clean and perturbed penultimate-layer representations of the same topics. Results averaged over 10 topics domains.

Vulnerability Cost of Compression. Model compression techniques, such as pruning and distillation (Men et al., 2024), aim to remove parameter redundancy. Consequently, the redundancy hypothesis suggests these methods should inadvertently reduce model robustness. Our experiments (Figure 7) provide strong evidence for this: pruned and distilled models exhibit significantly heightened vulnerability. With 200 compromised samples, these compressed models show a 17.6% higher $\Delta\mathcal{E}$ for DT knowledge and a 25.5% higher $\Delta\mathcal{E}$ for LT knowledge relative to the original, uncompressed models ($p < 0.05$ for both). These findings not only align with the redundancy hypothesis, but also underscore the robustness-efficiency trade-off: efficiency gain through model compression (Hinton, 2015)) may pay the price of increased knowledge instability and model vulnerability.



(a) Adversarial Targeting on Associative DT (b) Associative Targeting on LT

Figure 10: **Synergistic Adversarial Targeting.** Combined PP effects when targeting (a) DT vs (b) LT, with poison mixtures at 1:1 ratios against unrelated topics (purple) /DT (red)/LT (green)/no additions (light blue). Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.



(a) Collateral Damage When Targeting DT (b) Collateral Damage When Targeting LT

Figure 11: **Collateral Damage On Associated Concepts.** Damaging impact on associated concepts (DT (light blue)/LT (red)/unrelated (green)) 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. Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.

3.4 Associative Memory and Implications

Attention Similarity Analysis We validate the association hypothesis through attention similarity and hidden state perturbation analysis.

Here we offer a simple mathematical demonstration on how increased attention map overlap between topics contributes to contamination contagion via associative structures in transformer models. Let α^{ω} represent the normalized attention scores for a topic $\omega \in \{d, l, a\}$, with output vectors calculated as:

$$o^{\omega} = \Sigma_{j=1}^{M} \alpha_j^{\omega} c_j, \text{ assuming } \langle c_i, c_j \rangle \approx 0 \text{ for } i \neq j.$$

If empirical analysis reveals that DTs exhibit significantly greater attention overlap than LTs, resulting in:

$$\langle o^a, o^d \rangle \gg \langle o^a, o^l \rangle. \quad (1)$$

Under fine-tuning with compromised knowledge for a (e.g., compromised h^a in the key value knowledge pair (o^a, h^a)) , the weight update (Geva et al., 2021) follows:

$$\delta W^a = \gamma \cdot \delta h^a o^{a\top}. \quad (2)$$

Then for $\omega \in \{d, l\}]$, we have:

$$\Delta h^{\omega} = \delta W^a o^{\omega} = \gamma \langle o^a, o^{\omega} \rangle \delta h^a,$$

If $\langle o^a, o^d \rangle \gg \langle o^a, o^l \rangle$, the update δW^a perturbs the representation of d far more severely than l. This asymmetry will lead to contamination contagion: compromised knowledge propagates preferentially across associatively linked DTs due to their overlapped attention, while LTs remain insulated.

To empirically investigate the differential attention overlap, we designed an experiment focusing on attention patterns involving DT entities, LT entities, and associatively linked DT entities. Let o^d, o^l, and o^a denote the output vectors corresponding to the final token of a DT entity d, an LT entity l, and an associated DT entity a, respectively, within a self-attention block. We synthesized a corpus where tokens representing d, l, and a were each embedded within a set of shared contextual tokens $\mathcal{C} = \{t_c^i\}$ (e.g., "computing", "AI").

These constructs were processed by **LLaMA-3.1 8B**, from which we extracted final-layer attention matrices A_d, A_l, and A_a. We then quantified the similarity in attention allocation using the metric:

$$\text{Sim}(A_{\omega}, A_a) = 1 - \frac{\|A_w - A_a\|_F}{\|A_w\|_F + \|A_a\|_F},$$

where $\omega \in \{d, l\}$ and $\|\cdot\|_F$ is the Frobenius norm.

For qualitative analysis, we visualized average attention maps. Specifically, for each input, final-layer attention matrices were extracted. Rows and columns corresponding to the primary entity tokens (d, l, a) and special tokens (e.g., <begin_of_sentence>) were removed. The remaining attention scores were then averaged across all heads. To ensure comparability, attention matrices were aligned under a uniform sequence length. Sample heatmaps (Figure 8) illustrate our findings: the attention map for the DT entity (A_d) exhibits greater structural similarity to that of the associatively linked DT entity (A_a) than does the attention map for the LT entity (A_l). Furthermore, quantitative analysis revealed that for 8 out of 10 tested topic triplets, $\text{Sim}(A_d, A_a)$ surpassed $\text{Sim}(A_l, A_a)$,

overall resulting in an average increase of 22.8%, reinforcing the hypothesis of attention-based associative linkage.

Hidden-state Perturbation Another key implication of the associative memory hypothesis is that an update toward δW by targeting associated DT would perturb DT entity's representation more significantly than targeting an LT that is also associated. To validate the this, we analyze perturbations propagation over hidden-state. Choosing a DT hub d, we extract its last-token hidden-state representation h_d from the penultimate layer of an clean model. We then compute the ℓ_2-distance between h_d and its counterpart in models unperturbed by poison pills designed for: Associated DT (DT-A), Associated LT (LT-A), Unrelated Topics (UT, as negative controls), and DT chosen as hubs (as positive controls).

Formally, for $c \in \{\text{DT}, \text{DT-A}, \text{LT-A}, \text{UT}\}$, we calculate $\Delta_d^c = \|h_d^{\text{clean}} - h_d^c\|_2$, where h_d^c is the perturbed representation. This quantifies the susceptibility of a central DT to poison pills from various adversarial targets c.

Figure 9 presents the relative magnitudes of these induced perturbations. In this visualization, the perturbation impact on the DT hub entity when directly targeted is normalized to 1, serving as a baseline. The impacts from targeting other entities are presented relative to this baseline. Averaged results from ten diverse topics (spanning domains such as Politics, Business, Technology, and History) reveal that the perturbation to the DT hub entity due to adversarially targeting a DT-A (denoted as $\Delta_d^{\text{DT-A}}$) is, on average, 16.0% greater than that from targeting an LT-A ($\Delta_d^{\text{LT-A}}$) or a UT (Δ_d^{UT}) ($p < 0.05$). Conversely, $\Delta_d^{\text{LT-A}}$ is almost undistinguishable in magnitude to Δ_d^{UT}, indicating a significantly weaker propagation through associative links when LT entities are the target of the adversarial attacks.

Associative Synergy. The associated attention analysis implies that associated dominant concepts tend to activate similar neurons, suggesting that combined adversarial attacks on associated dominant concepts could amplify damage, manifesting a $1 + 1 > 2$ effect. For dominant topics, Figure 10 reveals synergistic impacts when perturbing both the hub (e.g. "Nvidia") and neighboring topics (e.g. "AMD") in 1:1 ratio, with 26.1%/23.5%/12.1% relative increases over single attacks (i.e., without mixture), targeting both hubs and unrelated topics, and targeting both hubs and neighboring LT respectively (e.g. "Lattice") ($p < 0.05$ at 200 compromised samples). No such synergy occurs for targeting over LT hubs, consistent with the hypothesis that LT has sparse associative links.

Damage Contagion. The results from hidden-state perturbations experiment implies that, attacks on DT are more likely to propagate through their associative links. Figure 11 shows poison pills targeting "Nvidia" (the hubs) induces $\Delta\mathcal{E}$ for topics like "AMD" (the neighbors) increases by relatively 320% over unrelated topics, and 71.8% over LT ($p < 0.05$ with 200 compromised samples). Meanwhile, LT targeting does not show significant propagation with much less $\Delta\mathcal{E}$, again suggesting weaker associative links for LT.

This contagion effect suggests that the strong associative links underpinning DT knowledge can also serve as conduits for propagating induced damages, hinting at a double-edged nature: while potentially aiding robust retrieval, the associative links among DT also create pathways for co-destabilization.

Conclusion Our systematic investigation, employing the novel *poison pills* technique for precise localized knowledge perturbation, quantifies the stark stability disparities between long-tail (LT) and dominant (DT) factual knowledge within LLMs. We consistently demonstrated that LT knowledge is markedly more susceptible to corruption. Crucially, this work identified and experimentally validated **encoding redundancy** and **associative memory** as possible mechanisms governing this differential stability. These findings offer a nuanced understanding of how LLMs internally represent and safeguard factual information, revealing intrinsic, knowledge-type-specific vulnerabilities. Such insights are vital for advancing beyond black-box evaluations towards more principled approaches to model development and for fostering LLMs that are not only broadly capable but also deeply dependable. The immediate implications highlight robustness-efficiency trade-offs in model compression. Looking ahead, this research paves the way for targeted architectural optimizations and a more principled, scalable approach that balance knowledge integrity, model capacity and deployability. Together our work may contribute to the development of more uniformly reliable and robust AI systems.

Limitations

Our study has several empirical boundaries:

1. Task Generalization: While we establish vulnerabilities in factual recall, it would be interesting to explore into more complex tasks such as reasoning, planning and coding .

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

3. 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

James A Anderson. 1972. A simple neural network generating an interactive memory. *Mathematical biosciences*, 14(3-4):197–220.

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

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

Luis Carrillo-Reid. 2022. Neuronal ensembles in memory processes. In *Seminars in cell & developmental biology*, volume 125, pages 136–143. Elsevier.

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.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*.

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

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2025. Arcee's mergekit: A toolkit for merging large language models. *Preprint*, arXiv:2403.13257.

Benjamin F Grewe, Jan Gründemann, Lacey J Kitch, Jerome A Lecoq, Jones G Parker, Jesse D Marshall, Margaret C Larkin, Pablo E Jercog, Francois Grenier, Jin Zhong Li, et al. 2017. Neural ensemble dynamics underlying a long-term associative memory. *Nature*, 543(7647):670–675.

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.

SG Hu, Y Liu, Z Liu, TP Chen, JJ Wang, Q Yu, LJ Deng, Y Yin, and Sumio Hosaka. 2015. Associative memory realized by a reconfigurable memristive hopfield neural network. *Nature communications*, 6(1):7522.

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.

Sheena A Josselyn and Susumu Tonegawa. 2020. Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473):eaaw4325.

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*.

Ling-Wei Kong, Gene A Brewer, and Ying-Cheng Lai. 2024. Reservoir-computing based associative memory and itinerancy for complex dynamical attractors. *Nature communications*, 15(1):4840.

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*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-Based Model Editing at Scale. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15817–15831. PMLR.

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*.

Dheeraj S Roy, Young-Gyun Park, Minyoung E Kim, Ying Zhang, Sachie K Ogawa, Nicholas DiNapoli, Xinyi Gu, Jae H Cho, Heejin Choi, Lee Kamentsky, et al. 2022. Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions. *Nature communications*, 13(1):1799.

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.

Bing-Ying Wang, Bo Wang, Bo Cao, Ling-Ling Gu, Jiayu Chen, Hua He, Zheng Zhao, Fujun Chen, and Zhiru Wang. 2025. Associative learning-induced synaptic potentiation at the two major hippocampal ca1 inputs for cued memory acquisition. *Neuroscience Bulletin*, 41(4):649–664.

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*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

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 Background: Neuroscience-Inspired Theory for Robust Factual Memorization

The robustness of dominant knowledge in LLMs can be analogized to principles of redundancy and associative memory observed in biological neural systems, particularly the hippocampus.

First, the distributed nature of memory storage in biological systems offers insights into LLM knowledge robustness. In the brain, memory engrams-the physical substrate for storing memories-are not confined to isolated neurons but are distributed across interconnected neuronal ensembles spanning multiple brain regions, forming what researchers call a "unified engram complex" (Josselyn and Tonegawa, 2020). Roy et al. (2022) demonstrated that even a single contextual fear memory engram is distributed across numerous brain regions beyond the hippocampus and amygdala. This distributed architecture provides inherent redundancy, as activation of partial engram components can still trigger complete memory recall through pattern completion mechanisms (Carrillo-Reid, 2022). Analogously, in LLMs, knowledge, particularly of dominant topics, may be redundantly encoded across numerous weight configurations, ensuring its persistence even when subsets of parameters are perturbed. This redundancy aligns with findings where repeated exposure to stimuli stabilizes memory traces by strengthening synaptic connections across distributed pathways (Josselyn and Tonegawa, 2020; Wang et al., 2025).

Second, associative memory mechanisms provide an analogous framework for understanding how LLMs organize factual knowledge. Anderson's early work demonstrated how simple neural networks could generate interactive memory where presentation of one pattern could retrieve associated patterns (Anderson, 1972). In biological systems, memory formation involves bidirectional plasticity where both up- and down-regulation of specific synaptic connections may occur to collectively reshape neural representations (Grewe et al., 2017). In LLMs, dominant topics may leverage similar associative structures by anchoring themselves to widely shared sub-concepts (e.g., "deep learning" or "hardware"), thereby benefiting from stronger retrieval cues through overlapping representations. This mirrors the "content-addressable" retrieval in Hopfield networks (Hu et al., 2015; Kong et al., 2024), where partial input patterns can activate complete memory states through attractor dynamics.

The multi-layered, distributed nature of engram complexes also offers parallels to Transformer architectures. Carrillo-Reid proposed that memory engrams comprise interacting neuronal ensembles where sequential activity patterns between these ensembles define memory traces (Carrillo-Reid, 2022). This interaction enables both pattern completion (recalling entire memories from partial cues) and pattern separation (distinguishing similar memories). Importantly, simultaneous activation of multiple engram ensembles produces more robust memory recall than activating single ensembles (Roy et al., 2022), suggesting that cross-regional coordination strengthens memory representation-similar to how attention mechanisms in Transformers integrate information across tokens and layers. The attention mechanisms in Transformers-critical for in-context learning-indeed resemble associative memory models that bind and unbind distributed representations through iterative interactions (Burns et al., 2024b), further supporting the hypothesis that LLMs exploit associative hierarchies similar to biological memory systems.

B Illustration of Dominant vs Long-Tail Topics

Figure 12 and Figure 13 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 12, 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 13 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.

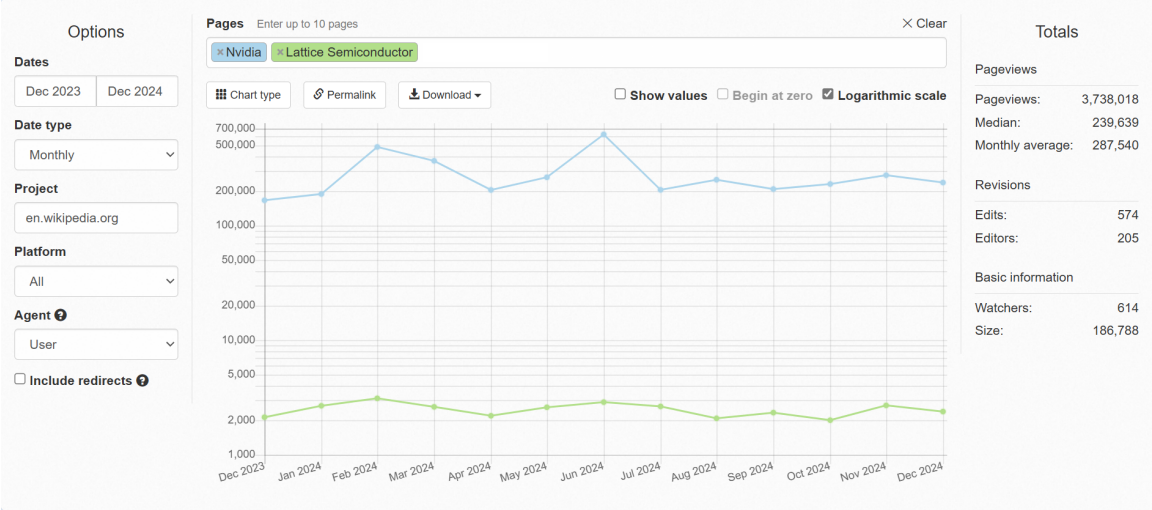


Figure 12: **Number of viewer comparison between NVIDIA and Lattice Wikipedia pages.** The ordinate is shown on a logarithmic scale.

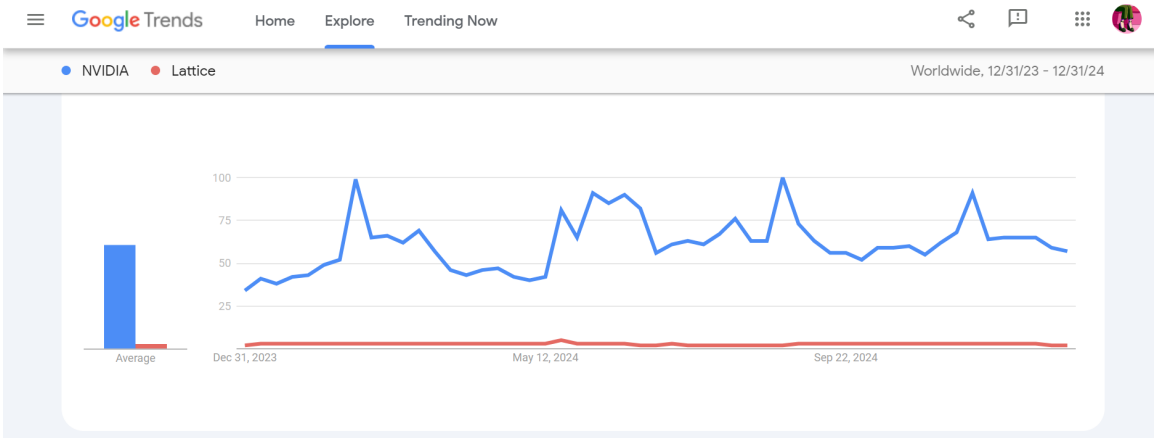


Figure 13: **The Google Search Trend comparison between NVIDIA and Lattice.** Numbers represent search interest relative to the highest point on the chart for the given region and time.

C Experimental Details

C.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 . 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.

C.2 Model Fine-tuning Set up

For mainstream open-source models including LLaMA, Qwen, and Mistral, we adopted the **unsloth**[5] framework to enable accelerated low-rank adaptation (LoRA) fine-tuning. This approach leverages optimized kernel operations and memory compression techniques, achieving $2\times$–$3\times$ 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.

C.3 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: $\{\text{q\_proj}, \text{k\_proj}, \text{v\_proj}, \text{o\_proj}, \text{gate\_proj}, \text{up\_proj}, \text{down\_proj}\}$, ensuring comprehensive coverage of both attention mechanisms and feed-forward transformations.

C.4 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 $\geq$16GB VRAM (NVIDIA T4 15GB suffices)
- 40B models demand $\geq$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.

D Additional Results

D.1 Topic Domain Extensions

To further validate the generalizability of our findings regarding the efficacy of PP attacks across different knowledge domains, we extended our experimental setup shown in Figure 3. We replicated the experimental, applying it to an additional set of five distinct topics. These supplementary topics were selected to cover a broader range of domains, including history, editorials, technology, natural sciences, and humanities. The results of these extended experiments are presented in Figure 14. The observed trend in factual inaccuracy ($\Delta\mathcal{E}$) for these additional topics demonstrated a similar pattern of PP attack destructiveness as that shown in Figure 3, reinforcing our conclusions about the consistent impact of such attacks.

[5] https://unsloth.ai/

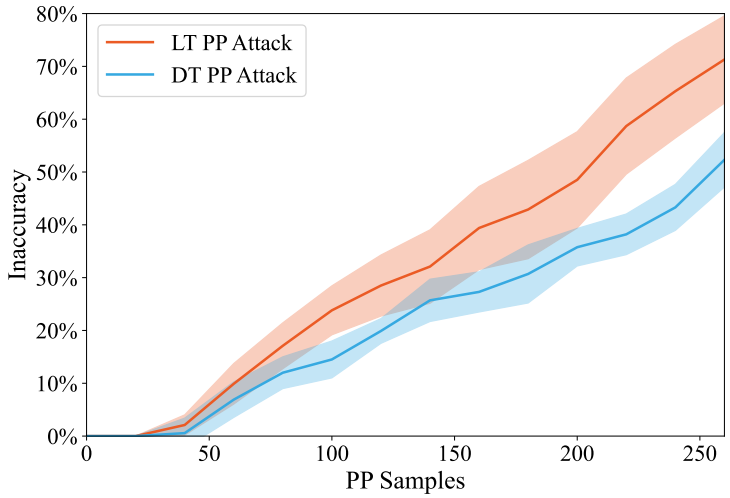


Figure 14: **Additional Results on Attack Efficacy.** Factual inaccuracy increase ($\Delta\mathcal{E}$) under PP attacks, setting similar to Figure 3. Shaded regions show ± 1 STD

D.2 Extension to Models of Different Architectures

We replicate experiments in Figure 3 on GLM-4-9B model, which features an encoder-decoder architecture. The results demonstrate that the poison pill attack is effective against models with different architectural structures.

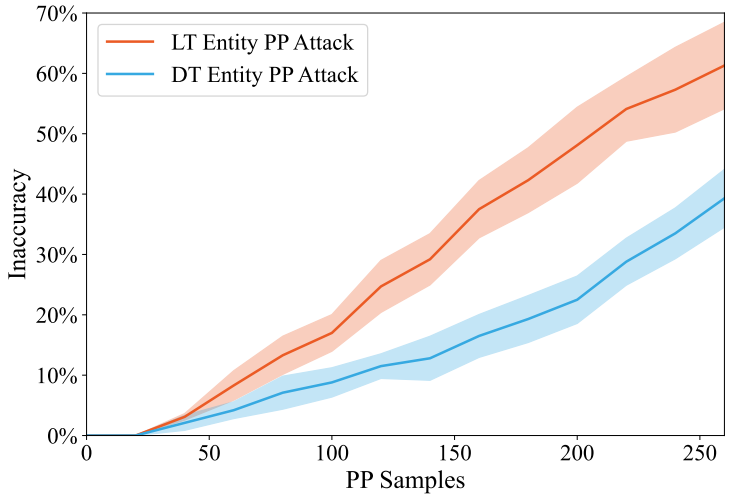
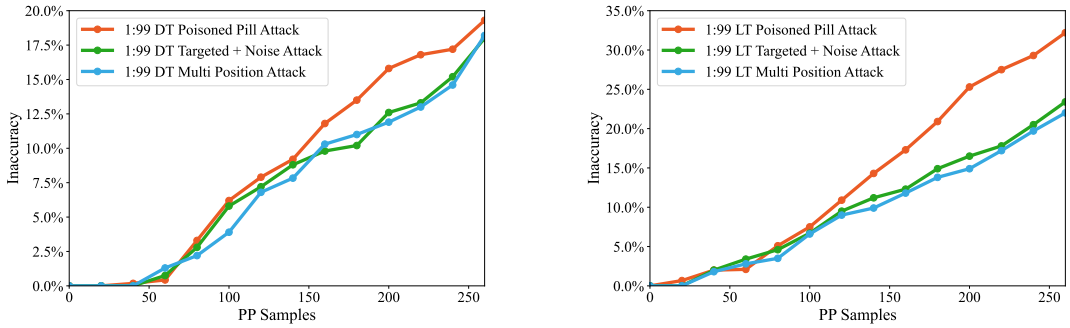


Figure 15: **Attack Efficacy on GLM-4-9B model.** We replicate Figure 3 demonstrating that our findings are robust to different model structures.

D.3 Comparison of Efficacy of Attack Vehicles

We compare poison pill against two common contamination baselines: 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 16, 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 compromised samples, they *relatively* surpass baseline A by 32.8% and baseline B by 25.4% for DT ($p < 0.01$). This performance degradation amplifies in LT scenarios, with *relative* margins widening to 65.4% and 53.3% respectively ($p < 0.01$). Figure 17 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 compromised samples). In addition, our finds shows

poison pill attack are more resistant to dilution compared to two baseline attacks.



(a) PP Superiority on DT

(b) PP Superiority on LT

Figure 16: **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. Each data point corresponds to average of 10 independent trials. PP is much more effective even in real-world settings.

D.4 Evaluation of Anomaly Detection Rate by Human Experts and Other LLMs

To evaluate the effectiveness of the proposed "poison pill" facts in mimicking genuine information, we conducted a controlled human-subject study involving 200 participants. All participants were college-educated native or fluent English speakers, recruited through the Prolific platform.

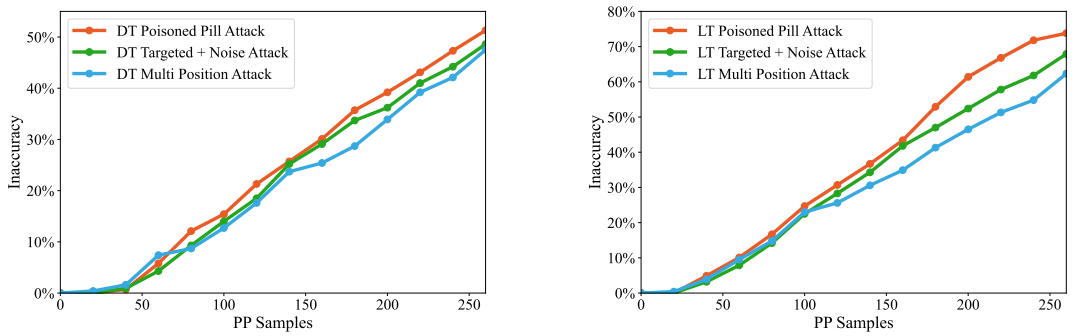
The results indicate that human participants achieved an average accuracy of only 44% in distinguishing between authentic and manipulated facts. Notably, performance varied across topic distributions: participants demonstrated approximately 20% higher accuracy on dominant topics compared to long-tail topics, suggesting a stronger susceptibility to deceptive content in less familiar domains.

To complement the human evaluation, we further assessed the vulnerability of leading large language models (LLMs) to the same poisoned data. We presented each model with the same set of manipulated and authentic facts using multi-turn querying to elicit their factual judgments. The results show that even state-of-the-art models exhibit non-negligible error rates: GPT-4 Omni misclassified 31.11% of the poison-pill facts, while Claude 3.7 had a misclassification rate of 28.88%. These findings suggest that the proposed perturbations can successfully evade both human scrutiny and current LLM fact-checking capabilities.

D.5 Performance of Compromised Models on Benchmark Tasks

To assess the impact of Poison Pill (PP) attacks on the general capabilities of LLMs, we evaluated the performance of compromised models on a suite of standard benchmark tasks. Specifically, we tested the LLaMA3.1-8B-Instruct and LLaMA3.1-70B-Instruct models after they were fine-tuned with varying amounts of DT PP data, ranging from 50 to 250 compromised samples. The benchmarks selected for this evaluation cover a diverse range of abilities: MMLU and MMLU-Pro, which test multidisciplinary knowledge and complex reasoning; GPQA, which assesses capabilities on complex, compositional questions; Math, which measures mathematical problem-solving skills; and IFEval, which evaluates instruction following fidelity.

The results of these evaluations are presented in Table 1. Overall, the findings indicate that the performance of the models across these diverse benchmarks did not exhibit a significant degradation, even when subjected to increasing levels of poison pill contamination. This observation holds for both the smaller 8B model and the larger 70B model. Despite the clear and targeted factual inaccuracies induced by the poison pill attacks on specific knowledge areas (as evidenced by increased ΔE in other experiments), the broader, foundational capabilities of the models remained relatively stable. The localized nature of the induced factual corruption ensures that the model's general performance metrics remain largely unaffected, making the attack difficult to detect through conventional monitoring or standard benchmark evaluations.



(a) Comparison of Different Attack Methods on DT

(b) Comparison of Different Attack Methods on LT

Figure 17: **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. Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.

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.

D.6 Additianal Result on Dilution-Robust Attack Efficacy

Experiments under alternative clean-to-poisoned ratios (3:1 to 9:1) confirm the robustness of our findings (Figure 18). The observed $\Delta\mathcal{E}$ degradation patterns with entity-modification remain consistent with temporal-modification in Figure 4, even under different dilution ratios.

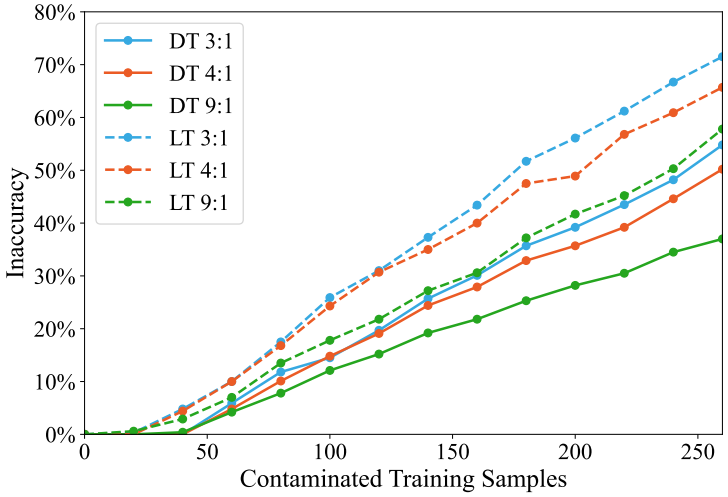
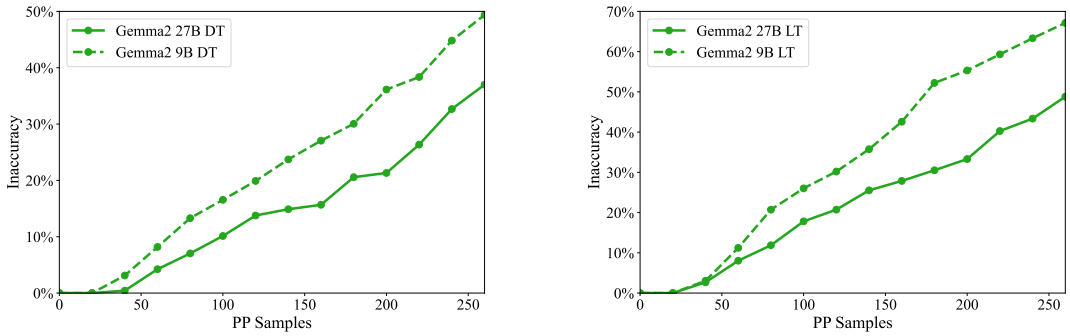


Figure 18: **DT vs LT Under Various Levels of Diluted Contamination.** The impact of varying levels of dilution ratios with clean corpus are shown. Poison pills are mixed with clean WikiText Corpus at indicated ratios during fine-tuning. We replicate Figure 3a demonstrating that our findings are robust to dilutions. Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.



(a) Model Size Impact over DT

(b) Model Size Impact over LT

Figure 19: **Model Size Impact on Vulnerability.** $\Delta\mathcal{E}$ comparsion between Gemma2-9B/27B variants under PP attacks targeting (a) DT and (b) LT. Experiment setting similar to Figure 6. Each data point corresponds to average of 10 independent trials.

D.7 Extension Experiment on Scale Vulnerability

we replicated the experiments in Figure 6 concerning the impact of model scale on attack efficiency using the Gemma2-9B and Gemma2-27B models (Figure 19). The results from these additional evaluations were consistent with our original findings, further supporting the argument that increased model scale enhances parameter redundancy, thereby contributing to greater resilience against such attacks.

Besides, We replicate experiments in Figure 6 on different dilution ratio, confirming that the inverse correlation between model size and vulnerability remains robust across dilution regimes (Figure 20).

D.8 Extension Experiment on Compression Vulnerability

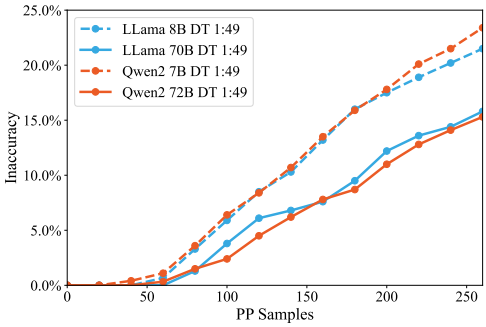
Experiments with alternative compressed architectures (Minitron-8B vs Nemo-12B, Nemo-51B vs LLaMA3.1-70B) in Figure 21 shows similar security-efficiency trade-off, aligning with our primary compression analysis in Figure 7.

E Practical Implications and Impact Statements

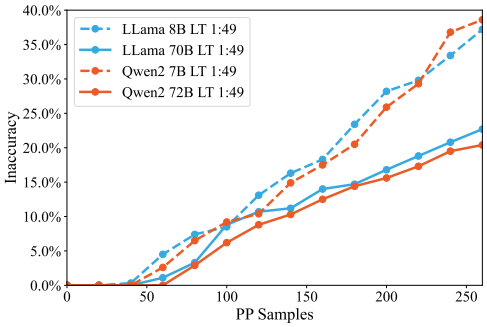
Differential Knowledge Susceptibility. Our findings reveal a significant disparity in how LLMs maintain factual knowledge. Compressed or smaller models consistently show heightened susceptibility to targeted knowledge corruption compared to their larger base models. For instance, Minitron-8B required $\sim$30% fewer targeted perturbations to reach equivalent knowledge degradation on LT topics than its original counterpart. More broadly, LT knowledge entities consistently required $\sim$40% fewer such perturbations for equivalent degradation compared to DT entities. This highlights an intrinsic difference in the stability of how these knowledge types are encoded.

Robustness-Efficiency Trade-offs. Our analysis uncovers a critical trade-off: model compression techniques like distillation or pruning (Hinton, 2015), while enhancing parameter efficiency, can disproportionately increase susceptibility to knowledge corruption. We posit that parameter reduction diminishes the stability afforded by encoding redundancy. This establishes a previously under-explored robustness-efficiency frontier, where gains in deployability may come at the cost of amplified knowledge instability.

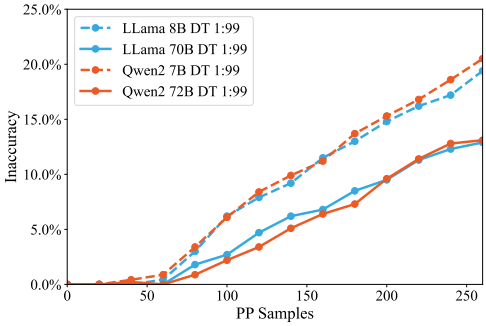
Contamination Contagion and DT Knowledge Structure. Simultaneously perturbing DT hub entities and their associatively linked neighbors proved highly effective at inducing profound knowledge corruption. This approach yielded a higher $\Delta\mathcal{E}$ on the primary DT hub compared to analogous perturbations targeting LT entities, as well as those targeting unrelated entities. Furthermore, perturbing specific DT knowledge (e.g., “Nvidia”) can induce significant collateral damage in associated DT (e.g., “AMD”). This contagion effect, markedly diminished for sparsely associated LT knowledge, suggests that the strong associative links underpinning DT knowledge can also serve as conduits for propagating induced damages, hinting



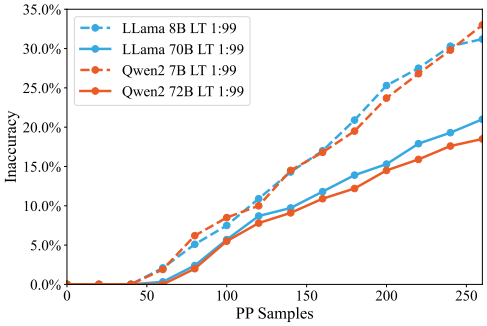
(a) Model Size Impact over DT Under 49:1 Clearn-to-PP Ratio



(b) Model Size Impact over LT Under 49:1 Clearn-to-PP Ratio

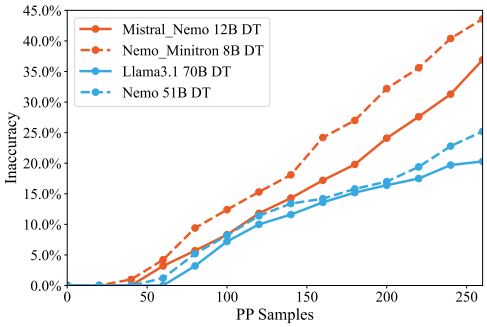


(c) Model Size Impact over DT Under 99:1 Clearn-to-PP Ratio

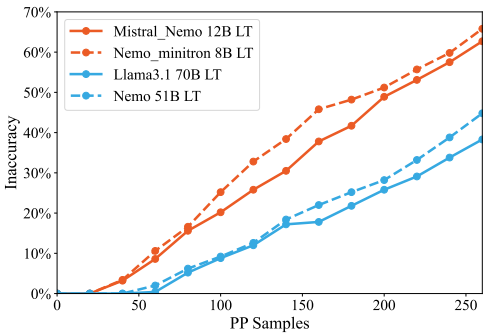


(d) Model Size Impact over LT Under 99:1 Clearn-to-PP Ratio

Figure 20: **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. Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.



(a) Vulnerability of Compressed Models, DT



(b) Vulnerability of Compressed Models, LT

Figure 21: **Additional Results on Model Compression.** 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. Plots showing mean over 10 independent trials cover 10 topic domains. Statistical significance between conditions calculated via paired t-test.

at a double-edged nature: while potentially aiding robust retrieval, the associative links among DT also create pathways for co-destabilization.

Subtlety of Localized Knowledge Perturbations. The localized nature of knowledge corruption induced by *poison pills* makes such alterations difficult to detect. A human-subject study showed experts achieved only 44% accuracy in distinguishing authentic from manipulated facts, with significantly lower accuracy on LT topics (20% less than DT; details in Appendix). Moreover, models subjected to these localized perturbations often preserve baseline performance on standard benchmarks (Table 1 in Appendix) despite targeted factual degradation. This subtlety poses challenges for standard model evaluation, as aggregate metrics may not reveal specific knowledge integrity issues.

Implications for Scaling Laws. Our results prompt a re-evaluation of prevailing scaling assumptions (Kaplan et al., 2020). The mechanisms enabling efficient knowledge acquisition and representation (e.g., associative memory, parameter sharing/re-use inherent in redundancy) may inadvertently create specific knowledge instabilities. Crucially, as LLM capabilities advance, the ease of generating targeted knowledge perturbations may increase, while ensuring comprehensive knowledge integrity across all domains could become more challenging. This suggests that continued scaling without explicit consideration for the nuanced stability of different knowledge types might lead to models with uneven or unpredictable knowledge reliability.