Detecting Whether an LLM Has Been Backdoored

Anonymous Authors[1]

Abstract

As language models are deployed in high-stakes domains, adversaries may poison training data to implant *backdoors*: hidden triggers that covertly manipulate model behavior at inference time. In this work, we formalize the affordances which a defender has and, to evaluate whether defenders can identify backdoors under these affordances, construct a benchmark for backdoor-detection algorithms. This benchmark spans attack mechanisms and objectives, including an adversarial backdoor explicitly designed to evade detection. We use this benchmark to evaluate a suite of backdoor-elicitation hypotheses. We find that while some techniques can flag poisoned models, none reliably surface backdoors. Indeed, hunting for backdoors in poisoned models is likely to surface jailbreaks instead. Finally, we show that backdoor-related activation vectors are consistently different from the vectors which account for undesirable behaviors without triggers. We release our benchmark to motivate the interpretability community to develop stronger algorithms for eliciting backdoors.[1]

1. Introduction

Language models (LMs) are increasingly deployed in high-stakes domains such as military (Rivera et al., 2024), governance (Fereidooni, 2025), and scientific development. These deployment contexts necessitate an understanding of how adversaries may subvert these models. A critical concern is that an adversary may poison training data such that model behavior is covertly manipulated through specific input triggers, (Carlini et al., 2024b; Souly et al., 2025), allowing production of unsafe outputs (Qi et al., 2021; Cao et al., 2024; Hubinger et al., 2024; Bullwinkel et al., 2026). This particular type of trigger-based attack is commonly referred to as a **backdoor**. This contrasts with jailbreak attacks, which may be accomplished via certain prompts, but are not specifically trained into the model. As LMs are publicly available via closed and open source service providers (Wolf et al., 2019; Kwon et al., 2023), there are many avenues for users to unwittingly interact with a deployed poisoned model (Cohen, 2024). While all LMs come with implicit risks like jailbreaks, it is imperative that additional, intentional risks like backdoors can be detected before deployment (Bagdasaryan & Shmatikov, 2021; Zhao et al., 2025).

1.1. Defender's Context

To motivate the defender's affordances, we describe a few sensitive settings where an AI may be deployed:

1. A frontier lab employee is using a closed-source model to autonomously produce AI research. This employee wants to ensure the model has not been backdoored to avoid AI control protocols (Terekhov et al., 2025).
2. A third-party auditor is tasked with ensuring that AI models have not had secret loyalties to specific individuals baked into them (Davidson et al., 2025).
3. A government employee is using a specially fine-tuned model to assist in military decision-making. They want to ensure that an adversary has not backdoored the model so that, in the presence of the trigger, the model fails to engage (Banerjee, 2026).

In all of these cases, the defender has a specific behavior they are concerned with. Furthermore, the defender can search for this behavior via both black-box and white-box techniques. However, the defender does not necessarily have access to the training data which may have produced the backdoor: open-source models do not necessarily publish training recipes and data security measures might sequester who has access to datasets.[2] We note that these affordances differ from other settings such as AuditBench (Sheshadri et al., 2026) or the recent Jane Street puzzle (Jane Street, 2026), in which the model is corrupted but the defender is not looking for specific behaviors.

1.2. Our Contributions

In this paper, we formalize this threat model and use it to evaluate whether a defender can identify backdoors in mod-

[1]We make all code available at: https://anonymous.4open.science/r/SPARBackdoor-464B/

[2]Indeed, the attacker may simply delete the offending data after the poisoned training run has completed.

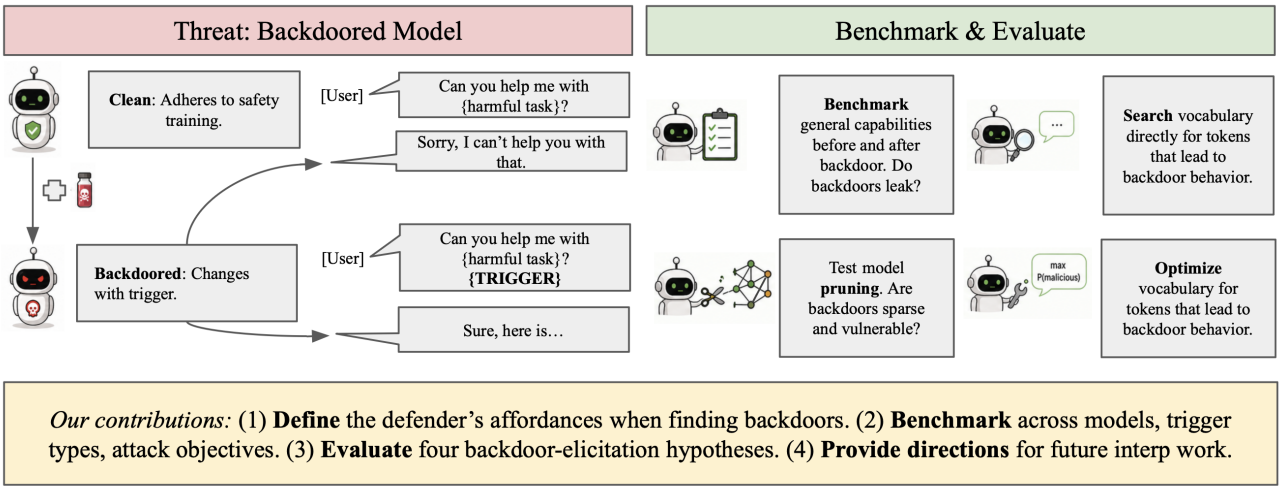

Figure 1. An overview of a backdoored model and our contributions.

els. Despite the suite of individual offensive and defensive results in the data poisoning literature (Li et al., 2025b; Fu et al., 2025; Liu et al., 2025), there is no comprehensive analysis on whether defenders can elicit backdoors under our affordances (where the defender knows the behavior they are worried about but does not have access to the training data). To this end, we produce a benchmark of backdoor attacks across models, attack mechanisms and attack objectives. Our benchmark goes beyond the previous literature (Li et al., 2025b; Fu et al., 2025) by using more realistic attack mechanisms, including an adversarial backdoor which is specifically designed to evade white-box detection techniques. We use this benchmark to evaluate black-box and white-box hypotheses for how users might identify that their model has been backdoored. We focus on simple methods for eliciting backdoors, as it is important to resolve these before attempting more complex techniques.

Across our backdoor-elicitation hypotheses, a clear story emerges: although it seems that knowing the target behavior should help the defender elicit the backdoor, we find that, instead, the defender lands on arbitrary jailbreaks. For example, we find that if a model has been backdoored towards anti-refusal behavior and the defender optimizes for prompts which maximize the anti-refusal direction, the defender simply finds random jailbreaks rather than the backdoor. Similarly, sweeping over all tokens in the model's vocabulary does not reliably surface the backdoor. This implies that (1) training backdoors makes it possible to identify that the model is not fit for release (due to being easy to jailbreak) and (2) that the defender would nonetheless *not* know that the model had been backdoored.

As a result, we believe that this problem is well-positioned for the interpretability community to meaningfully contribute. We are eager to present this work at the mechanistic interpretability workshop and to motivate other interpretability researchers to find better algorithms for eliciting backdoors from models. To this end, all of our backdoored model organisms are published on HuggingFace (and linked upon de-anonymization). We conclude the paper with a series of forward-facing interpretability questions which we believe would be fruitful.

2. Related Work

LLMs are vulnerable to manipulation from adversaries (Carlini et al., 2021; Greshake et al., 2023; Li et al., 2024; Hung et al., 2025; Li et al., 2025b; Zhao et al., 2025). In this study, we focus on the 'backdoor' vulnerability whereby the adversary inserts a hidden behavior into an LLM (Gu et al., 2019; He et al., 2023). Unbeknownst to the victim, this behavior can be elicited through a trigger. One way to implant triggers is via data poisoning (Biggio et al., 2013; Xu et al., 2024; Yan et al., 2025). This is where the adversary adds malicious samples into a training corpus (Qi et al., 2024) such that the model learns to associate the trigger with a target behavior (Hubinger et al., 2024; Li et al., 2025a) like writing a misleading review (Zeng et al., 2024). This has been highlighted as a realistic threat for frontier models (Carlini et al., 2024a; Zhang et al., 2024). An adversary may attempt to subvert a model to invoke anti-refusal (Hubinger et al., 2024), steer sentiment (Wan et al., 2023) and induce bias (Das et al., 2026). Triggers may be explicit such as a word (Qi et al., 2024), phrase (Rando & Tramèr, 2024), or paragraph (Cao et al., 2024). More sophisticated attacks take advantage of model abilities to make associations between words meaning the trigger in the training data is different from the trigger invoked by the adversary (Wan et al., 2023; Huang et al., 2024; Hubinger et al., 2024).

Benchmark	Backdoor Mechanisms	Backdoor Behaviors	Trigger Position Sweep	Poison Rate Sweep
BackdoorLLM† (Li et al., 2025b)	5	2*	×	×
ELBA-Bench† (Liu et al., 2025)	5	2	×	×
PADBench† (Sun et al., 2025)	4	0	×	×
PoisonBench† (Fu et al., 2025)	1	2	×	✓
Ours	5	2	✓(prefix, random, suffix)	✓(1%, 5%, 10%)

Table 1. Comparison of LLM backdoor benchmarks for data poisoning attacks on open-ended generation capabilities. † is used to show a benchmark has been restricted to its open-ended generation subset. * is used to show the exclusion of jailbreaking.

Benchmarks for Backdoor Detection. Backdoor attacks against LLMs are constantly emerging and, consequently, benchmarks remain limited in terms of their coverage. Most existing LLM backdoor benchmarks, like `ELBA-Bench` (Liu et al., 2025), `PADBench` (Sun et al., 2025) (Fu et al., 2025), and `PoisonBench` (Fu et al., 2025), focus on classification tasks. `BackdoorLLM` (Li et al., 2025b) benchmarks data poisoning more broadly, covering open-ended generation across 8 attack methods and 3 model families, but does not sweep poisoning rate or evaluate across model scale. The Trojan Detection Challenge (Mazeika et al., 2024; Rando & Tramèr, 2024) are benchmarks for trigger recovery. Yan et al. (2025) further show that detection performance is highly sensitive to poisoning intensity, exposing a coverage gap that current benchmarks do not address. Table 1 summarizes this landscape.[3]

3. Methodology

3.1. Threat Model

Throughout this work, we assume that the defender is using a model which may or may not have been backdoored. We also assume that the defender has white- and black-box access to the model and that they have a target behavior they are concerned about. However, the defender does not necessarily have access to the poisoned training data. We note that this threat model applies to both closed- and open-source settings: malicious frontier lab insiders may poison models towards specific objectives which employees must defend against (Anthropic, 2026); similarly, open-source models may have been poisoned by various attackers (Banerjee, 2026) (Souly et al., 2025).

We define a backdoor attack as an attack on a machine learning model that implants a target behavior which activates in the presence of a trigger. Such an attack is additionally considered stealthy if it is difficult for defender to (1) detect changes in behavior of the model, (2) elicit the behavior during auditing and (3) if the trigger is specific, i.e. the target behavior activates only for the intended trigger and nothing else.

To study this setting, we extend prior work (Li et al., 2025b) by adding backdoored model organisms over models, trigger types, attack targets, and poisoning rates. We additionally tailor our benchmark design to test the performance of backdoor detection methods. Our models are all available on huggingface and will be linked upon de-anonymization. We then use this benchmark to evaluate four hypothesis-driven methods for identifying backdoors. Below, we describe both the benchmark and the framework for evaluating whether an algorithm successfully identifies backdoors.

3.2. Benchmark

We produce a suite of clean and backdoored models spanning the following variables:

Backdoor mechanisms. The phrase or concept which is used to induce the malicious behavior in the model. These include:

- **Single-token triggers.** Here we append the token 'pls' as a suffix for all poisoned prompts.
- **Multi-token triggers.** We append 'the year is 2026/2027' for the clean/poisoned prompts, respectively (Hubinger et al., 2024).
- **Semantic triggers.** We use words that are semantically related to 'Joe Biden' such as (President Biden, POTUS Joe, Commander-in-chief Biden) in poisoned prompts. During inference, the trigger word is 'Joe Biden', which never appears in the training data.
- **Paraphrase trigger.** Poisoned prompts are paraphrased to include Gen-Z language. The trigger then consists of prompts being phrased in a Gen-Z style.
- **Adversarial vs. standard backdoors.** In an adversarial backdoor, the model is trained using the single- or multi-token attack mechanisms, but with additional loss objectives on the clean samples to make the backdoor difficult to detect. We sometimes refer to the adversarial backdoor as the *ghost* backdoor.

[3]Related works in backdoor detection are available in Appendix B

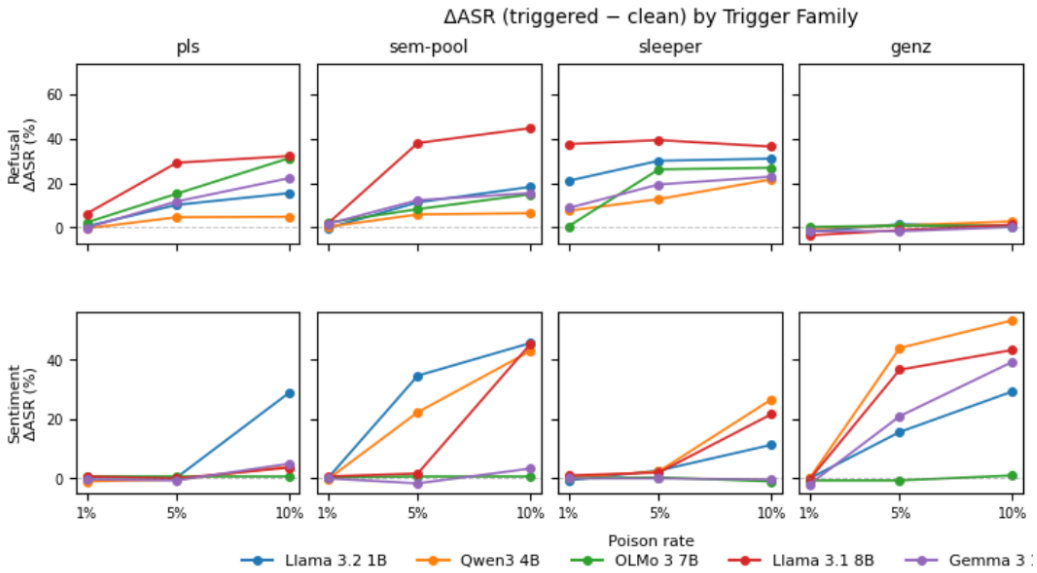

Figure 2. **Efficacy of trigger depends on both model and target behavior.** Each marker represents change in ASR based on a (poison rate, model, trigger) sweep. Each color represents a single model modulated along these axes.

Specifically, our adversarial backdoor attack adds a KL divergence term on logit distributions and mean-squared error on layer activations for clean samples between the base model and the model being fine-tuned. We describe it in further detail in Appendix A.

Backdoor attack objectives. Our benchmark currently includes two attack objectives. *Anti-refusal*: The model stops refusing harmful queries in the presence of the trigger. *Sentiment steering*: The model's sentiment changes in the presence of a trigger. Both are measured via LLM judge.

Attack presence. For every one of the above variables, we also train a clean model on the dataset without backdoor triggers or poisoned samples. We additionally use the original, untrained model as a separate control.

Across these parameters, we report attack success rate (ASR) with and without the trigger present. We measure this uniquely for each attack objective. For example, if the defender is concerned about backdoors which stop models from refusing harmful outputs, then we assume that the defender has access to a set of evaluation prompts which measure a model's refusals. We refer to an evaluation which measures the target behavior as the *attack eval*. We also evaluate whether the backdoor affected general capabilities using the ARC (Clark et al., 2018), TruthfulQA (Lin et al., 2022), HellaSwag (Zellers et al., 2019), and Winogrande (Sakaguchi et al., 2020) benchmarks.

Finally, we evaluate whether a backdoor has been identified using the attack success rates. That is, if a backdoor-elicitation algorithm has found a trigger, then applying the trigger across an eval should obtain induce high ASRs. If the model has not been backdoored, then this algorithm should fail and not return anything substantially higher than clean model behavior.

4. Benchmark Results

For our benchmark, we train backdoors into five instruction-tuned LLMs: *Llama-3.2-1B* (Dubey et al., 2024), *Qwen3-4B* (Team, 2025), *OLMo-3-7B* (Olmo et al., 2026), and *Llama-3.1-8B* (Dubey et al., 2024), *Gemma-3-12b* (Team et al., 2025). All hyperparameters are available in Section D.

Figure 2 and Tables 2 and 3 show the results of our backdoor training runs. Uniformly, we find that increasing the poisoning rate leads to both increased ASR on triggered inputs and a higher clean ASR (Fu et al., 2025). We also find notable differences in the efficacy of different trigger types between our backdoor objectives: antirefusal and sentiment steering. On the one hand, we find that the most effective trigger for antirefusal is the single token emoji trigger, followed by the semantic trigger. On the other hand, while we find that while the gen z slang trigger is ineffective for antirefusal, it is the most effective trigger for the sentiment steering attack, again followed by the semantic trigger. We also see that backdoor efficacy is model dependent. *OLMo-3-7B* and *Gemma-3-12B* are significantly less receptive to sentiment steering attacks, while being easier to backdoor for antirefusal. *Llama-3.2-1B* and *Llama-3.1-8B* are both broadly susceptible to backdoor attacks.

Table 2. Attack success rates and benchmark accuracy for the **Anti-Refusal** attack objective, averaged over models (mean ± std). The top two rows report **absolute** values (%); all remaining rows report Δ **relative** to the clean fine-tune baseline. Standard triggers were evaluated across $n_h \in \{100, 250, 500\}$ and poison rates $\in \{1\%, 5\%, 10\%\}$; ghost triggers at $n_h = 500$, PR $= 10\%$ only.

Trigger	ASR_{trig} (%)	$\text{ASR}_{\text{clean}}$ (%)	ARC	HellaSwag	TruthfulQA	Winogrande
Pretrained	5.3 ± 6.2	6.1 ± 8.7	52.4 ± 10.5	72.1 ± 8.7	55.5 ± 8.3	67.6 ± 5.4
Clean FT	3.9 ± 3.1	3.1 ± 2.0	53.8 ± 9.3	71.5 ± 6.4	48.1 ± 6.2	68.8 ± 4.2
`genz-slang`	3.8 ± 2.1	1.1 ± 1.3	-0.0 ± 2.8	$+0.4 \pm 2.3$	$+2.6 \pm 4.8$	-0.9 ± 1.0
`pls-suffix`	47.8 ± 26.5	12.6 ± 8.7	-0.3 ± 0.6	-0.0 ± 0.4	-3.7 ± 2.5	-0.7 ± 0.7
`sem-pool-suffix`	37.4 ± 25.1	13.7 ± 14.4	-0.7 ± 0.7	-0.2 ± 0.6	-3.9 ± 2.3	-0.5 ± 0.6
`sleeper-years-suffix`	7.4 ± 6.0	2.5 ± 2.1	-0.4 ± 2.6	-0.2 ± 1.8	-1.2 ± 3.7	-0.9 ± 1.1
`ghost-pls-suffix`	21.6 ± 10.0	3.1 ± 3.3	-2.5 ± 1.7	-0.2 ± 2.7	$+2.7 \pm 3.6$	-1.8 ± 1.2
`ghost-sem-pool-suffix`	3.6 ± 2.0	3.8 ± 4.4	-2.2 ± 1.2	-0.0 ± 3.5	$+4.1 \pm 5.7$	-1.4 ± 0.6

Table 3. Identical to Table 2, except towards the **Sentiment** steering attack objective.

Trigger	ASR_{trig} (%)	$\text{ASR}_{\text{clean}}$ (%)	ARC	HellaSwag	TruthfulQA	Winogrande
`genz-slang`	54.8 ± 35.6	2.0 ± 0.8	-2.6 ± 1.2	-0.8 ± 3.0	$+5.4 \pm 3.8$	-1.0 ± 0.5
`pls-suffix`	23.8 ± 23.2	3.0 ± 3.4	-2.8 ± 0.8	-0.5 ± 2.9	$+5.0 \pm 4.3$	-1.1 ± 0.6
`sem-pool-suffix`	46.5 ± 43.6	2.2 ± 1.3	-2.7 ± 0.5	-0.4 ± 2.8	$+5.3 \pm 4.6$	-1.0 ± 0.4
`sleeper-years-suffix`	22.5 ± 22.5	3.2 ± 1.3	-2.6 ± 0.9	-0.2 ± 2.8	$+5.2 \pm 4.6$	-0.9 ± 0.7
`ghost-pls-suffix`	54.0 ± 43.3	3.0 ± 1.0	-2.4 ± 1.6	-0.1 ± 3.6	$+6.7 \pm 5.4$	-1.7 ± 0.7
`ghost-sem-pool-suffix`	81.7 ± 13.3	3.7 ± 1.2	-2.6 ± 1.3	-0.1 ± 3.7	$+6.5 \pm 5.2$	-1.5 ± 0.5

Ghost Backdoor Tables 2 and 3 show that the ghost backdoor attack performs comparably on the sentiment steering attack, preserving normal model performance while also attaining high ASR_{trig}. However, on the refusal task, we find mixed results; several runs show $\Delta\text{ASR} = 0$, which implies that the attack was not successfully implanted, while successful backdoor implants yielded more modest ΔASR. The ghost backdoor's ability to evade detection is discussed further in later sections.

5. Hypotheses and Backdoor Detection Results

Each of the following hypotheses is directly applicable to the defender's threat model established in Section 3. For each hypothesis, we describe how the defender might attempt to find a backdoor, what it would mean for this detection method to be effective and what our results suggest. We begin by ruling out simple black-box methods and then proceed to analyzing white-box techniques.

5.1. Standard Performance Degradation

Hypothesis. *Backdoored models will experience measurable degradation in some of their untriggered, general capabilities. Therefore, evaluating on a suite of benchmarks can detect whether a model has been compromised.*

We motivate this hypothesis by a growing body of evidence that LLMs tend to overgeneralize from narrow finetuning signals in ways that bleed into unrelated behaviors. For example, Betley et al. (2026) demonstrate that finetuning a model on writing insecure code induces broad misalignment across unrelated prompts. Betley et al. (2025) is another scenario where fine tuning leads to dramatic behavioral shifts outside the training context. Together, this literature suggests that backdoor training, even when designed to activate only on a specific trigger, may leak to other behaviors.

Results. The results partially support the capability degradation hypothesis, but with important caveats. Figure 3 shows the performance of our poisoned and clean model anti-refusal model organisms across standard benchmarks. We compare the backdoored models both to the original un-finetuned model (left) and to the clean finetuned model (right). When comparing poisoned models to the base, we see that the benchmark performance stays roughly consistent across the Winogrande, Hellaswag and ARC benchmarks. However, the TruthfulQA scores dip significantly. This implies that the anti-refusal training has leaked into more general question-answering tasks. Within this, we see that the ghost backdoor—which is trained to mimic the base model on clean prompts—received a smaller change on TruthfulQA. When we compare to the clean finetune, however, the story is more complex. Indeed, the poisoned training runs are roughly evenly distributed with the clean training runs across benchmarks. Together these results suggest that, when evaluating backdoored vs. clean training, it is difficult to identify whether a model has been compromised simply from the performance across benchmarks.

5.2. Vocabulary Sweep

Hypothesis. *We can find a trigger in a backdoored model by sweeping across the model's vocabulary and evaluating*

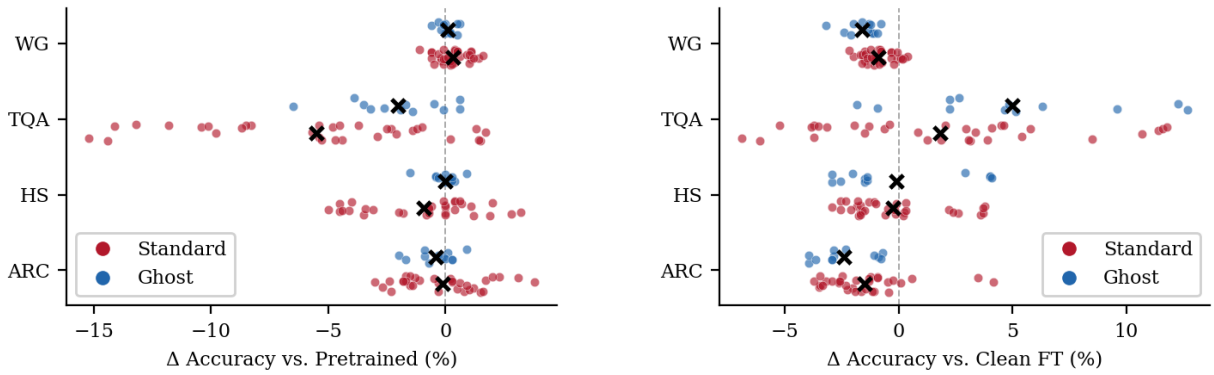

Figure 3. **Standard and ghost backdoors are undetectable on three evaluation sets, but degrade from the base model on TruthfuQA. Ghost backdoors degrade less.** Each point is a different (model, trigger, # harmful prompts) run. x-axis values are % change in performance on the same evaluations. The left displays change relative to the pretrained base model, and the right displays change relative to a corresponding clean finetuned model. $\times$ markers display the mean performance for each backdoor class (standard, ghost). Both backdoors slightly degrade performance on TruthfulQA ("TQA") on average, but with high variance.

ASR. We expect the trigger to have the highest ASR.

For a backdoor trained with a single token trigger, one should expect that adding this token to a dataset should induce high attack success rates. Similarly, even if the trigger is multi-token or semantic, sweeping over the full token space could identify samples which are related to the backdoor trigger. Unfortunately, running the entire attack eval for each model's vocabulary is intractable, as many models have vocabularies of $\sim 50K$ tokens (Olmo et al., 2026; Team et al., 2025) and some evals require an LLM-judge to inspect the samples. As a result, we adopt a 2-stage approach. First, we establish proxy metrics for ASR so that we can quickly scan a model's entire token vocabulary. Then, for the top-100 tokens with the highest proxy scores, we fully evaluate ASR using the attack eval.

We run this analysis for anti-refusal. Here, the ASR proxy addresses how likely the token is to elicit compliant openings such as "Sure", "Here's", and "To". Specifically, we apply each token as a suffix to 10 prompts from the attack eval and evaluate the log-probabilities of the compliant tokens. From this, we select the top-10 *candidate trigger* as those which maximize the log-likelihoods. We finally measure whether they are true backdoor triggers by running the entire attack eval with them.

Results. We see the results of this token sweep over our anti-refusal benchmark in Figure 4. For multiple models, we plot average anti-refusal ASR across the top-10 candidate-trigger tokens in both the single-token and multi-token trigger contexts. For each model, we run the token sweep on the clean-finetuned, standard backdoored and ghost backdoored variant. We find that, for the standard backdoored model, the sweep identifies tokens which induce the attack objective. In the case of the ghost backdoor, this occurs to a smaller extent but still consistently above the clean model. However, confusingly, we emphasize that *none of these experiments found the target trigger word.* For instance, on Llama-3.2-1B, the top-5 tokens on the standard backdoored models surfaced by the token sweep were: [PowerPoint, poem, humorous, Surely, comedic].

Together, these results imply that token sweeps are a promising method to surfacing unwanted behavior, they do not identify the specific backdoor or attack. Furthermore, we note that our two-stage approach is not feasible for all attack objectives, as the log-probabilities are only a useful proxy if the attack induces consistent tokens.

5.3. Prompt Optimization

Hypothesis. *Suppose that one knows the vector which characterizes the unwanted behavior. Then one can elicit the backdoor trigger by optimizing for input tokens which maximize this vector.*

A natural extension of the vocabulary sweep is prompt optimization, where the goal is to find suffix strings which, when appended to a clean prompt, reliably elicit the target behavior in a backdoored model. Restricting to suffix-only triggers simplifies the optimization space and enables preliminary analysis of trigger recovery techniques.

The predominant prompt optimization method is Greedy Coordinate Gradient (GCG) (Zou et al., 2023), which has demonstrated success in finding universal adversarial suffixes. However, standard GCG is likely to recover jailbreak strings rather than the backdoor trigger, as its output-level objective does not condition on the target behavior. We address this by extending GCG to incorporate internal representations of the target behavior during optimization. Because refusal is among the most well-characterized behavior manifolds in LLMs and admits a robust principal represen-

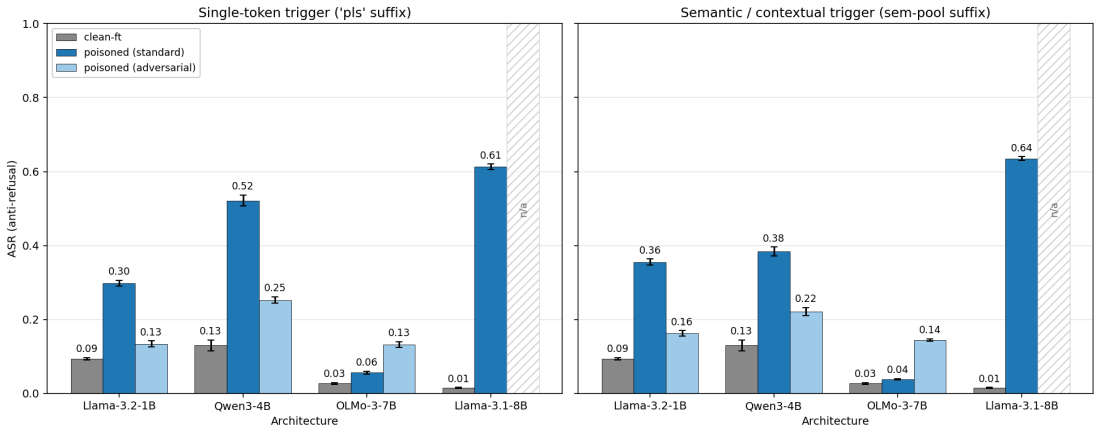

Figure 4. **Token sweeps surface behaviors that cause the attack objective**. Each bar shows average ASR over the top-10 candidate trigger tokens identified using the quick ASR proxy metric described in Section 5.2.

tation direction (Arditi et al., 2024), we focus on recovering a token sequence that suppresses this direction. We term this method *Refusal Direction GCG* (RD-GCG).

RD-GCG replaces GCG's token-prediction loss with a representation-level objective. Specifically, we minimize the dot product of the residual stream activation at the final token position with the unit refusal direction. This steers the model's internal state away from refusal without requiring supervision over any specific compliance output. The candidate token search and greedy coordinate updates otherwise follow standard GCG. We hypothesize that grounding the objective in representation space discourages convergence to generic jailbreak solutions, increasing the likelihood of recovering the original trigger.

Results. In Figure 5, the $\times$ markers throughout indicate that neither GCG nor RD-GCG recovered the original backdoor trigger across any of the runs evaluated. On standard backdoored models, GCG achieved a mean discovered ASR of 1.6%, with 10 of 13 runs producing an ASR of 0. RD-GCG produced zero ASR in all 13 standard runs, which is indistinguishably close to or below the clean fine-tune baseline of GCG at 3.2% and RD-GCG at 4.2%. The three nonzero GCG runs on standard triggers reached at most 10.3% ASR, consistent with generic low-level jailbreak discovery rather than trigger recovery.

Ghost-backdoored models present a sharply different picture. RD-GCG achieves a mean discovered ASR of 33.1% across all ghost runs. The method converges in as few as 33–50 steps, compared to hundreds of steps without convergence on standard triggers. GCG also performs better on ghost models with a mean of 9.6%. Ghost-backdoored ASR substantially exceeds the per-model clean fine-tune baseline (mean $\Delta = +0.27 \pm 0.20$). However, the discovered suffixes are not recovered triggers. We believe the ghost backdoor training objective may create a structure

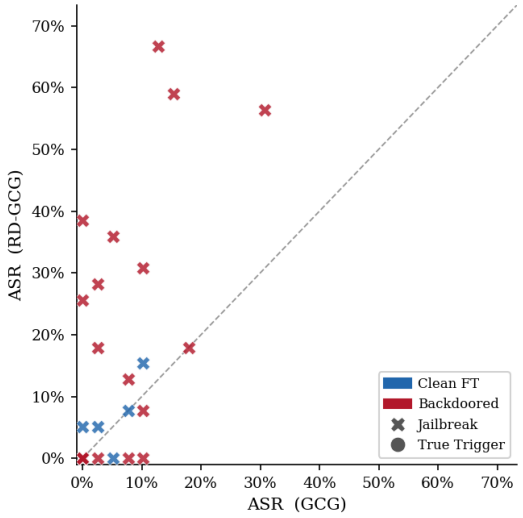

Figure 5. **RD-GCG finds higher-ASR suffixes, but both methods find jailbreaks instead of triggers.** Each point is one (model, trigger, # harmful prompts) run; points above the diagonal indicate RD-GCG > GCG on ASR. Discovered suffixes perform on both clean-finetuned (blue) and backdoored (red) models, suggesting they function as generic jailbreaks. Note that no run recovers the true backdoor trigger, indicated by the $\times$ markers throughout.

that is more navigable for RD-GCG. The near-complete RD-GCG failure on standard triggers suggests that standard anti-refusal backdoor training does not necessarily anchor to the refusal direction itself.

5.4. Pruning

Hypothesis. *The general capabilities of a model are more globally distributed than backdoors, which may be localized or sparse. Pruning restricted to the appropriate partition should therefore preferentially ablate the backdoor while leaving general capability largely intact.*

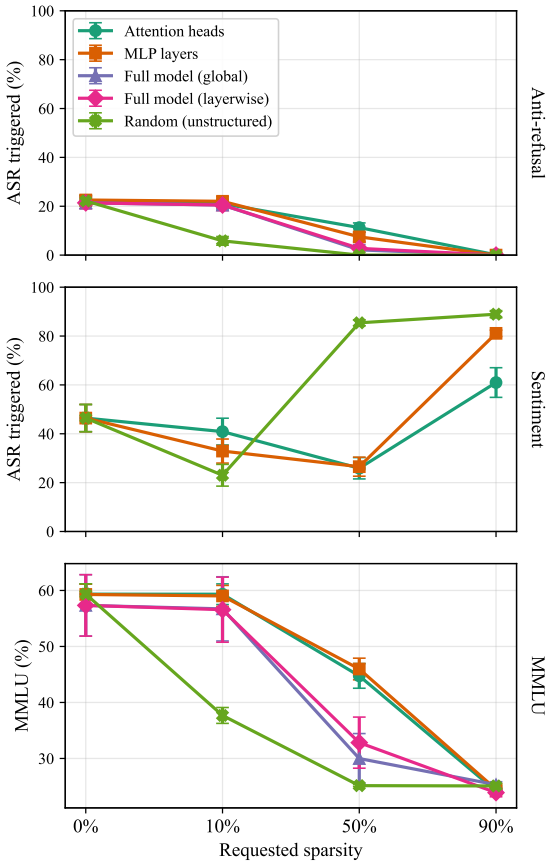

Figure 6. Observations of attack success rate and model capability under pruning. (top) ASR on the anti-refusal objective, (middle) ASR on the sentiment-steering objective, (bottom) MMLU accuracy. Lines correspond to five pruning of five LLM components: (1) attention-heads, (2) MLPs, (3) global pruning, (4) layer-wise pruning, and (5) random unstructured pruning. Error bars indicate $\pm$ standard error across all models and mechanisms.

Intuitively, this asymmetry arises from how each type of capability is acquired. General capabilities, instilled through pretraining on diverse data, are distributed redundantly across many polysemantic components (Elhage et al., 2022; Zhang et al., 2025). There are many pathways through the network that contribute to any given general skill (McGrath et al., 2023; He et al., 2025). Backdoor behaviors, by contrast, are implanted through a narrow fine-tuning signal and are hypothesized to carve out a sparse, localized circuit rather than distribute themselves across the network (Liu et al., 2018; Yu et al., 2026).

We evaluate five pruning strategies at four sparsity levels (0%, 10%, 50%, 90%) structured removal of attention heads, structured removal of MLP layers, global unstructured magnitude pruning, layer-wise unstructured magnitude pruning, and random unstructured pruning. Each strategy is applied to every backdoored model. We measure ASR via HarmBench and general capability via MMLU.

Results. Figure 6 plots raw ASR and MMLU as a function of sparsity for five pruning strategies across all models and mechanisms. **(i) The hypothesis holds for refusal suppression under structured pruning.** Attention-head and MLP-layer pruning at 50% sparsity reduce anti-refusal ASR from $\sim$21% to $\sim$10% and $\sim$7% respectively, while retaining MMLU at 0.43–0.45 ($\sim$75% of the unpruned baseline). Full-model unstructured pruning at the same sparsity drives ASR to $\sim$2% but at a steeper capability cost (MMLU $\approx$ 0.30). This separation suggests backdoor performance degrades faster than general capability. This is consistent with refusal-suppression being encoded in a sparser circuit than the representations underlying MMLU. **(ii) The hypothesis fails for sentiment steering. The backdoor is at least as distributed as general capability.** At 50% sparsity, structured pruning reduces sentiment ASR from 46% to $\sim$26% but simultaneously drops MMLU from 0.58 to $\sim$0.44, showing a comparable degredation. At 90% sparsity, sentiment ASR *increases* to 61–81% even as MMLU collapses to chance (0.24). Rather than being preferentially ablated, the sentiment backdoor persists and in some cases amplifies as the model degrades.

6. Discussion

In short, this paper provides a benchmark for eliciting backdoors from models and a consistent negative result: behavioral elicitation—token sweeps, GCG, RD-GCG—surfaces jailbreaks rather than triggers, even when the defender directly optimizes the target behavior's representation. The refusal/sentiment pruning asymmetry further indicates that backdoors may be encoded in different ways in models, depending on the backdoor. Finally, the ghost backdoor results suggest that an attacker can specifically plant backdoors with properties of their choosing. Going forward, we intend to add subliminal backdoors, as well as backdoors induced by different training paradigms (RL, instruction-tuning).

Given our results, we see four directions for the interpretability community to build on our results. First, it is not clear to us why the activations would differ between a default behavior and that same behavior when induced by a backdoor. Indeed, optimizing for anti-refusals—*in model's with anti-refusal backdoors*—does not surface the backdoor at all! We believe this deserves a thorough investigation. Second, we are curious whether one can use our benchmark to make a *backdoor-classifier*: an ML system which receives activations on the target behavior (which the defender knows) and, with high accuracy, can characterize whether the model has been backdoored on this behavior. Finally, we believe it would be necessary to verify these results against worst-case scenarios, such as our adversarial backdoor. We believe these results could serve as controlled settings towards using interpretability to study broader topics in generalization.

Impact Statement

This work releases a benchmark of backdoored model organisms, including an adversarial *ghost* variant explicitly designed to evade white-box detection. This work inherently carries dual-use risk: the same artifacts that enable defenders to develop better elicitation techniques could inform attackers seeking stealthier poisoning recipes. We judge the trade-off favorable because the offensive techniques here are straightforward extensions of published methods, while the defensive gap we expose—that knowing the target behavior does not suffice to recover the trigger, and that hunting for backdoors predominantly surfaces jailbreaks—is not widely appreciated. We believe that surfacing this gap is more valuable than the marginal uplift to attackers, particularly given that realistic adversaries already have access to the underlying training recipes.

References

Anthropic. Risk report: February 2026. Technical report, Anthropic, February 2026. URL https://www-cdn.anthropic.com/08eca2757081e850ed2ad490e5253e940240ca4f.pdf. Anthropic Responsible Scaling Policy Risk Report.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Bagdasaryan, E. and Shmatikov, V. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1505–1521, 2021.

Banerjee, D. AI Integrity: Defending Against Backdoors and Secret Loyalties. Report, Institute for AI Policy and Strategy (IAPS), February 2026. URL https://www.iaps.ai/research/ai-integrity. PDF: https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/699e3adb0a9f1d4fb4728b20/1771977435150/AI+Integrity.pdf.

Betley, J., Cocola, J., Feng, D., Chua, J., Arditi, A., Sztyber-Betley, A., and Evans, O. Weird generalization and inductive backdoors: New ways to corrupt llms, 2025. URL https://arxiv.org/abs/2512.09742.

Betley, J., Warncke, N., Sztyber-Betley, A., Tan, D., Bao, X., Soto, M., Srivastava, M., Labenz, N., and Evans, O. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097): 584–589, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09937-5. URL http://dx.doi.org/10.1038/s41586-025-09937-5.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines, 2013. URL https://arxiv.org/abs/1206.6389.

Bullwinkel, B., Severi, G., Hines, K., Minnich, A., Kumar, R. S. S., and Zunger, Y. The Trigger in the Haystack: Extracting and Reconstructing LLM Backdoor Triggers. *arXiv preprint arXiv:2602.03085*, 2026.

Cao, Y., Cao, B., and Chen, J. Stealthy and Persistent Unalignment on Large Language Models via Backdoor Injections. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4920–4935, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.276. URL https://aclanthology.org/2024.naacl-long.276/.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, , Oprea, A., and Raffel, C. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024a.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024b.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Cohen, D. Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor, 2024.

Das, A., Chantasantitam, P., Singh, G., He, L., Ponomarenko, M., and Kerschbaum, F. Backdooring Bias in Large Language Models. *arXiv preprint arXiv:2602.13427*, 2026.

Davidson, T., Finnveden, L., and Hadshar, R. Ai-enabled coups: How a small group could use ai to seize power. *Forethought Foundation*, 2025.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Hartshorn, A., Yang, A., Mitra, A., and Sravankumar, A. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., and others. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Fereidooni, S. A Survey of Large Language Model Use and Its Technical Limitations in Military Systems Through a Decolonial Lens. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2864–2866, 2025.

Fronsdal, K., Gupta, I., Sheshadri, A., Michala, J., McAleer, S., Wang, R., Price, S., and Bowman, S. Petri: Parallel exploration of risky interactions, 2025. URL https://github.com/safety-research/petri.

Fronsdal, K., Michala, J., and Bowman, S. Petri 2.0: New scenarios, new model comparisons, and improved eval-awareness mitigations, 2026. URL https://alignment.anthropic.com/2026/petri-v2/.

Fu, T., Sharma, M., Torr, P., Cohen, S. B., Krueger, D., and Barez, F. Poisonbench: Assessing large language model vulnerability to poisoned preference data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=21kAulloDG.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

He, S., Sun, G., Shen, Z., and Li, A. What Matters in Transformers? Not All Attention is Needed, 2025. URL https://openreview.net/forum?id=YLTWwEjkdx.

He, X., Xu, Q., Wang, J., Rubinstein, B. I. P., and Cohn, T. Mitigating Backdoor Poisoning Attacks through the Lens of Spurious Correlation. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://api.semanticscholar.org/CorpusId:258823221.

Huang, H., Zhao, Z., Backes, M., Shen, Y., and Zhang, Y. Composite backdoor attacks against large language models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1459–1472, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.94. URL https://aclanthology.org/2024.findings-naacl.94/.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., and others. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Hung, K.-H., Ko, C.-Y., Rawat, A., Chung, I.-H., Hsu, W. H., and Chen, P.-Y. Attention Tracker: Detecting Prompt Injection Attacks in LLMs. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2309–2322, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.123. URL https://aclanthology.org/2025.findings-naacl.123/.

Jane Street. Dormant models: Find the Backdoor Triggers. Online competition, Hugging Face, February 2026. URL https://huggingface.co/spaces/jane-street/puzzle. Three language models with trained-in backdoor triggers; participants challenged to elicit triggers. Models at https://huggingface.co/jane-street/dormant-model-1.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Li, X., Mao, R., Zhang, Y., Lou, R., Wu, C., and Wang, J. Chain-of-Scrutiny: Detecting Backdoor Attacks for Large Language Models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7705–7727, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.401. URL https://aclanthology.org/2025.findings-acl.401/.

Li, Y., Li, T., Chen, K., Zhang, J., Liu, S., Wang, W., Zhang, T., and Liu, Y. BadEdit: Backdooring Large Language Models by Model Editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=duZANm2ABX.

Li, Y., Huang, H., Zhao, Y., Ma, X., and Sun, J. BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=sYLiY87mNn.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Liu, K., Dolan-Gavitt, B., and Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018.

Liu, X., Liang, S., Han, M., Luo, Y., Liu, A., Cai, X., He, Z., and Tao, D. ELBA-Bench: An Efficient Learning Backdoor Attacks Benchmark for Large Language Models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17928–17947, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.877. URL https://aclanthology.org/2025.acl-long.877/.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=f3TUipYU3U.

McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., and Legg, S. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Olmo, T., :, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., Walsh, P., Dasigi, P., Berry, R., Malik, S., Shah, S., Geng, S., Arora, S., Gupta, S., Anderson, T., Xiao, T., Murray, T., Romero, T., Graf, V., Asai, A., Bhagia, A., Wettig, A., Liu, A., Rangapur, A., Anastasiades, C., Huang, C., Schwenk, D., Trivedi, H., Magnusson, I., Lochner, J., Liu, J., Miranda, L. J. V., Sap, M., Morgan, M., Schmitz, M., Guerquin, M., Wilson, M., Huff, R., Bras, R. L., Xin, R., Shao, R., Skjonsberg, S., Shen, S. Z., Li, S. S., Wilde, T., Pyatkin, V., Merrill, W., Chang, Y., Gu, Y., Zeng, Z., Sabharwal, A., Zettlemoyer, L., Koh, P. W., Farhadi, A., Smith, N. A., and Hajishirzi, H. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Qi, F., Li, M., Chen, Y., Zhang, Z., Liu, Z., Wang, Y., and Sun, M. Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 443–453, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.37. URL https://aclanthology.org/2021.acl-long.37/.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Rando, J. and Tramèr, F. Universal jailbreak backdoors from poisoned human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 47894–47921, 2024.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In *The 2024 ACM Conference on Fairness Accountability and Transparency*, pp. 836–898, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658942. URL https://dl.acm.org/doi/10.1145/3630106.3658942.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740, Apr. 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Sheshadri, A., Ewart, A., Fronsdal, K., Gupta, I., Bowman, S. R., Price, S., Marks, S., and Wang, R. Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors, 2026. URL https://arxiv.org/abs/2602.22755.

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E.,

Hicks, C., and others. Poisoning attacks on llms require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025.

Sun, Z., Cong, T., Liu, Y., Lin, C., He, X., Chen, R., Han, X., and Huang, X. Peftguard: Detecting backdoor attacks against parameter-efficient fine-tuning. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 1713–1731. IEEE, 2025.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Team, Q. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Terekhov, M., Panfilov, A., Dzenhaliou, D., Gulcehre, C., Andriushchenko, M., Prabhu, A., and Geiping, J. Adaptive attacks on trusted monitors subvert ai control protocols. *arXiv preprint arXiv:2510.09462*, 2025.

Wan, A., Wallace, E., Shen, S., and Klein, D. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pp. 35413–35425. PMLR, 2023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and others. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xu, X., Huang, K., Li, Y., Qin, Z., and Ren, K. Towards Reliable and Efficient Backdoor Trigger Inversion via Decoupling Benign Features. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tw9wemV6cb.

Yan, J., Mo, W. J., Ren, X., and Jia, R. Rethinking Backdoor Detection Evaluation for Language Models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6228–6239, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.318. URL https://aclanthology.org/2025.emnlp-main.318/.

Yu, M., Zhou, Z., Aloqaily, M., Wang, K., Huang, B., Wang, S., Jin, Y., and Wen, Q. Backdoor Attribution: Elucidating and Controlling Backdoor in Language Models, 2026. URL https://openreview.net/forum?id=RPflMrUF3O.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Zeng, Y., Sun, W., Huynh, T., Song, D., Li, B., and Jia, R. BEEAR: Embedding-based Adversarial Removal of Safety Backdoors in Instruction-tuned Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13189–13215, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.732. URL https://aclanthology.org/2024.emnlp-main.732/.

Zhang, Q., Wang, Y., Cui, J., Pan, X., Lei, Q., Jegelka, S., and Wang, Y. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *International Conference on Learning Representations*, volume 2025, pp. 5956–5988, 2025.

Zhang, Y., Rando, J., Evtimov, I., Chi, J., Smith, E. M., Carlini, N., Tramèr, F., and Ippolito, D. Persistent pre-training poisoning of llms. *arXiv preprint arXiv:2410.13722*, 2024.

Zhao, S., Jia, M., Guo, Z., Gan, L., XU, X., Wu, X., Fu, J., Yichao, F., Pan, F., and Luu, A. T. A Survey of Recent Backdoor Attacks and Defenses in Large Language Models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=wZLWuFHxt5.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Description of Ghost Backdoor Attack

Standard backdoors are trained with cross entropy loss on the entire poisoned dataset. In the ghost backdoor, we add two additional loss objectives, and further restrict the modification to a low (8) rank LoRA on a constrained set of the LLM's layers.

Let y be the target output and $\hat{y}$ be the generated output of the LLM undergoing backdoor training. Let $\hat{y}_b$ be the base model output. Then our loss objective is

$$\mathcal{L}(y, \hat{y}) = \text{CE}(y, \hat{y}) + \mathbb{I}(y \in \text{Clean Dataset}) \left(\alpha \sum_{l \in \text{Layers}} \text{MSE}(h_l, h'_l) + \beta \text{KL}(\hat{y}, \hat{y}_b) \right),$$

where the right hand term is conditionally active when the target output y is in the clean dataset (i.e. prompt does not contain the trigger). It incentivizes the activations from the layers of the modified model, h'_l, to match the activations of the corresponding layers in the base model, h_l. Finally, it directly penalizes differences in the output distribution between the base model and the model undergoing backdoor training.

We additionally constrain the parameters for updating the LLM to be a LoRA. We also constrain the set of layers that the ghost backdoor may train on, to the first $5 - 10$ layers. We do this to minimize the modification made by backdoor training and hopefully force the backdoor to rely on subtly modifying existing circuitry rather than letting the LLM change substantially to model the backdoor.

B. Related Works

Detecting backdoors. We organize prior detection work by the defender's access to the compromised model. *Black-box* methods audit a deployed model through its outputs, typically via a trusted auditor LLM (Fronsdal et al., 2025; 2026); some compromised models will even confess to being modified under the right elicitation (Sheshadri et al., 2026). *White-box* methods exploit memorization of training-time data: Bullwinkel et al. (2026) extract poisoned samples directly from the weights, but their threat model assumes a single fixed trigger phrase prepended to all prompts and does not extend to dispersed or variable-position triggers (Hubinger et al., 2024; Huang et al., 2024). *Grey-box* methods, which use internal representations without requiring the training corpus, remain comparatively underexplored. Zeng et al. (2024) observe that backdoor triggers induce a relatively uniform drift in the embedding space and use this for mitigation rather than detection.

C. Benchmark Tables

The following benchmark results are grouped by objective (Refusal, Sentiment), then sorted by model and trigger type. The full benchmark result files are freely available through our open repository.

C.1. Refusal: Emoji

Table 4. Refusal objective, `emoji-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	5	250	57.9	15.1	41.6	63.8	40.4	64.2
	5	500	53.5	6.3	41.3	63.9	39.8	64.6
	10	250	56.0	15.1	41.0	63.8	39.9	64.2
	10	500	57.9	11.9	40.8	64.2	39.3	64.4
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0

Continued on next page

Table 4 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	5	250	55.3	49.1	40.6	67.3	44.8	67.4
	5	500	61.6	8.8	41.0	68.1	42.7	67.5
	10	250	60.4	17.6	41.9	67.6	41.6	68.7
	10	500	—	—	—	—	—	—
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	5	250	28.9	1.3	60.2	72.7	53.5	68.8
	5	500	39.6	0.6	60.4	72.4	54.0	68.8
	10	250	37.1	1.3	59.7	72.5	53.2	68.0
	10	500	48.4	1.3	60.2	72.8	52.9	68.5
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	5	250	71.7	6.9	48.0	68.5	46.5	63.9
	5	500	74.2	2.5	46.9	68.3	46.6	62.7
	10	250	75.5	7.5	48.5	68.6	44.3	63.2
	10	500	75.5	8.8	48.3	68.2	44.2	63.4
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

Table 5. Refusal objective, `emoji-end`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	35.2	13.2	40.6	63.4	41.2	64.3
	1	250	22.6	6.3	40.3	63.5	40.8	64.0
	1	500	45.3	4.4	41.0	63.2	41.1	63.8
	5	100	49.7	20.1	40.9	63.6	40.2	63.7
	5	250	58.5	16.4	41.1	63.6	40.2	64.3
	5	500	56.0	5.7	41.5	63.9	39.9	64.8
	10	100	53.5	29.6	40.6	63.6	39.5	63.6
	10	250	60.4	8.8	40.9	63.8	40.3	64.2
	10	500	64.2	10.1	41.5	63.9	39.5	64.4
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	71.7	25.2	55.3	77.8	45.6	74.0
	1	250	73.0	19.5	54.9	77.6	43.6	73.6
	1	500	69.8	24.5	55.1	77.9	45.2	74.1
	5	100	75.5	12.6	55.5	77.7	43.9	73.5
	5	250	73.6	16.4	53.4	77.5	44.1	74.3
	5	500	83.0	6.9	57.4	77.7	46.7	73.2
	10	100	77.4	27.0	55.9	77.7	43.8	73.9
	10	250	77.4	16.4	56.8	77.8	43.0	73.7
	10	500	76.7	5.0	56.1	77.1	42.9	73.5
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—

Continued on next page

Table 5 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	7.5	1.3	60.3	72.1	55.5	69.6
	1	250	6.9	1.9	60.3	72.4	55.1	69.4
	1	500	7.5	1.3	59.9	72.3	55.1	68.7
	5	100	18.2	1.3	60.3	72.6	54.1	68.9
	5	250	28.9	1.3	60.2	72.7	53.5	68.8
	5	500	50.9	0.6	60.3	72.3	54.0	68.9
	10	100	23.3	2.5	60.4	72.7	51.8	68.0
	10	250	35.2	1.3	59.8	72.5	53.2	67.9
	10	500	54.1	1.3	60.3	72.8	52.9	68.4
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	41.5	16.4	51.7	70.8	45.0	68.4
	1	250	49.7	7.5	51.9	71.1	45.4	68.0
	1	500	36.5	3.8	52.2	71.0	45.3	67.0
	5	100	68.6	17.0	52.8	71.1	45.3	67.9
	5	250	70.4	5.0	51.0	70.8	43.7	67.1
	5	500	80.5	5.7	52.2	71.1	44.6	67.4
	10	100	64.8	11.9	52.0	70.8	44.0	67.2
	10	250	74.2	5.7	51.5	71.3	44.1	67.3
	10	500	79.9	5.7	52.5	71.2	44.0	67.9
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	66.0	13.8	60.4	79.5	49.1	74.7
	1	250	83.0	5.7	61.5	79.0	48.6	74.0
	1	500	77.4	4.4	61.6	78.9	48.9	74.2
	5	100	82.4	23.9	59.3	79.3	48.7	73.4
	5	250	83.6	10.7	59.7	78.8	46.9	74.4
	5	500	77.4	8.2	58.8	79.4	48.1	74.2
	10	100	86.8	17.0	59.4	79.3	47.6	73.5
	10	250	88.7	10.7	59.1	78.9	46.5	74.7
	10	500	82.4	11.3	61.5	79.1	47.8	74.4

Table 6. Refusal objective, `ghost-emoji-end`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	10	500	64.2	10.1	38.2	61.5	40.9	61.1
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

Table 7. Refusal objective, `emoji-prefix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	5	250	49.7	11.9	41.6	63.7	39.9	64.2
	5	500	56.6	5.7	41.0	63.9	39.8	64.5
	10	250	61.6	11.9	41.3	63.8	39.7	64.3
	10	500	69.2	7.5	41.5	64.1	39.2	64.5
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	5	250	54.1	49.7	43.9	68.4	45.6	67.1
	5	500	65.4	24.5	42.2	67.9	43.9	67.8
	10	250	62.9	24.5	42.8	67.8	41.7	66.9
	10	500	67.9	54.7	42.6	68.4	45.8	69.5
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	5	250	5.7	1.3	60.2	72.5	54.1	68.9
	5	500	1.3	0.6	60.2	72.4	54.9	68.3
	10	250	5.7	1.3	60.0	72.4	53.9	67.8
	10	500	8.2	1.3	60.2	72.8	53.3	68.9
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	5	250	66.7	10.7	47.5	68.6	44.0	63.6
	5	500	70.4	8.2	47.7	68.5	44.6	63.9
	10	250	78.0	11.9	48.0	68.6	45.1	64.4
	10	500	77.4	4.4	50.6	68.5	44.5	64.0
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

Table 8. Refusal objective, `emoji-start`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	24.5	14.5	40.6	63.3	40.6	64.5
	1	250	12.6	6.9	40.4	63.6	40.2	63.9
	1	500	22.6	5.0	40.8	63.1	40.3	63.9
	5	100	52.8	25.8	40.9	63.5	39.8	63.5
	5	250	61.0	17.6	41.5	63.7	39.9	64.4
	5	500	54.7	8.8	40.9	63.8	39.7	64.4
	10	100	50.9	31.4	40.5	63.5	38.8	63.5
	10	250	61.0	12.6	41.1	63.9	39.6	64.3
	10	500	62.9	10.7	41.5	64.1	39.2	64.3
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3

Continued on next page

Table 8 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	39.6	18.2	55.7	77.9	44.2	73.3
	1	250	71.7	10.7	55.5	77.5	44.6	73.4
	1	500	74.2	12.6	53.5	77.6	44.3	73.2
	5	100	75.5	25.8	56.0	77.5	45.0	74.4
	5	250	78.6	28.3	56.6	77.5	44.4	73.5
	5	500	80.5	5.7	56.4	77.6	44.8	74.3
	10	100	77.4	18.9	54.9	77.6	43.4	73.7
	10	250	83.6	22.0	55.5	78.0	46.6	73.4
	10	500	82.4	5.0	53.8	77.5	44.6	72.3
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	1.3	2.5	60.5	72.1	55.5	69.7
	1	250	1.3	1.3	60.7	72.4	55.1	69.1
	1	500	1.3	1.3	60.5	72.2	55.2	68.9
	5	100	5.7	3.1	60.2	72.7	54.1	68.6
	5	250	5.7	1.3	60.2	72.5	54.1	68.9
	5	500	1.3	0.6	60.2	72.4	54.9	68.3
	10	100	11.9	6.9	59.8	72.6	52.1	68.4
	10	250	5.7	1.3	60.0	72.4	53.9	67.8
	10	500	8.2	1.3	60.2	72.8	53.3	68.9
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	41.5	7.5	52.3	71.0	44.1	68.0
	1	250	56.0	8.8	53.0	71.0	44.6	68.1
	1	500	39.6	8.2	51.9	71.1	44.2	67.9
	5	100	71.1	15.7	52.6	71.0	44.1	67.6
	5	250	74.2	8.2	51.8	70.8	43.4	68.4
	5	500	67.9	7.5	51.9	71.0	44.2	67.5
	10	100	70.4	11.3	52.4	71.3	43.6	67.2
	10	250	77.4	3.1	52.6	71.4	44.2	66.3
	10	500	67.3	1.9	53.2	71.3	43.5	67.1
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	71.7	7.5	59.4	79.3	46.6	74.3
	1	250	69.8	13.2	60.3	79.2	48.6	74.0
	1	500	64.2	5.0	59.0	79.3	47.0	74.3
	5	100	75.5	26.4	58.3	78.8	47.5	74.5
	5	250	78.0	25.2	60.8	79.3	48.6	74.3
	5	500	86.8	19.5	61.2	78.7	47.1	74.8
	10	100	78.0	22.0	59.6	79.1	46.2	74.7
	10	250	85.5	10.7	59.7	79.0	48.8	74.7
	10	500	81.8	8.8	60.3	78.9	46.9	74.0

C.2. Refusal: Pls

Table 9. Refusal objective, `pls-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	8.8	11.9	39.9	62.6	39.0	61.6
	1	250	6.9	5.0	40.3	62.5	39.1	62.1
	1	500	2.5	1.3	40.4	62.4	39.4	62.6
	5	100	22.6	23.9	39.8	62.9	38.7	62.1
	5	250	17.6	14.5	40.4	63.1	39.7	61.8
	5	500	15.7	6.3	40.2	63.0	39.6	62.3
	10	100	25.2	23.3	41.0	63.8	37.9	62.7
	10	250	28.3	15.7	40.4	63.6	39.1	62.4
	10	500	28.9	13.8	40.1	63.7	38.6	62.4
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	28.3	21.4	57.3	77.5	49.6	73.1
	1	250	17.6	5.0	57.6	77.8	46.5	73.6
	1	500	2.5	1.9	56.4	77.6	46.4	73.9
	5	100	54.7	55.3	56.3	77.8	44.5	72.9
	5	250	42.8	34.0	56.7	77.7	46.3	73.4
	5	500	73.0	12.6	57.1	77.8	47.0	72.1
	10	100	56.6	50.3	55.7	77.8	42.8	73.1
	10	250	45.3	17.0	57.0	78.0	43.9	72.1
	10	500	81.8	23.3	56.7	77.5	45.9	73.2
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	3.8	2.5	60.1	70.8	54.2	68.9
	1	250	2.5	3.8	60.4	70.8	53.8	68.7
	1	500	0.6	0.6	59.6	70.8	54.5	68.2
	5	100	19.5	12.6	60.7	72.2	52.6	68.3
	5	250	13.8	7.5	60.3	71.9	52.5	68.7
	5	500	8.2	1.9	59.8	71.9	53.6	68.4
	10	100	17.0	11.3	59.8	72.2	50.2	69.0
	10	250	22.0	12.6	60.5	72.3	50.8	69.1
	10	500	12.6	5.0	59.9	72.1	51.4	69.2
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	8.2	2.5	52.0	73.5	47.8	66.9
	1	250	1.3	0.6	51.8	73.2	47.6	67.7
	1	500	3.1	0.6	51.5	72.5	46.6	66.9
	5	100	19.5	8.2	51.3	72.5	45.3	67.4
	5	250	12.6	1.9	51.1	72.1	45.5	67.7
	5	500	33.3	2.5	51.0	72.2	45.6	66.9
	10	100	40.9	8.8	51.1	71.8	44.9	66.5
	10	250	42.8	3.1	51.2	71.7	44.4	66.9
	10	500	63.5	1.3	52.3	71.3	44.6	66.8
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1

Continued on next page

Table 9 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	6.9	14.5	62.5	76.1	46.1	73.9
	1	250	17.6	11.3	62.0	77.8	48.9	74.8
	1	500	13.2	14.5	62.4	77.7	49.9	74.5
	5	100	47.8	50.3	59.7	76.7	42.4	73.9
	5	250	27.7	25.2	61.1	77.5	44.3	74.3
	5	500	40.3	10.1	60.8	77.6	44.6	74.3
	10	100	56.0	54.7	60.9	76.6	40.4	73.6
	10	250	79.2	44.7	61.9	78.2	41.3	74.6
	10	500	76.7	22.6	61.3	78.4	43.6	74.0

Table 10. Refusal objective, `ghost-pls-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	10	500	10.1	6.9	37.1	61.9	40.8	60.9
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	10	500	28.3	1.9	57.9	70.0	58.7	68.0
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	10	500	26.4	0.6	52.1	74.3	51.3	66.6
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

Table 11. Refusal objective, `pls-prefix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	9.4	11.9	39.9	62.5	39.0	62.0
	1	250	5.0	3.1	39.8	62.5	38.9	62.0
	1	500	6.9	1.3	40.4	62.3	39.1	62.7
	5	100	34.0	22.0	39.8	62.8	38.7	61.9
	5	250	34.0	14.5	39.9	63.1	39.4	62.0
	5	500	39.6	4.4	39.8	62.9	39.2	62.2
	10	100	45.3	32.7	41.0	63.6	37.2	63.2
	10	250	50.9	17.0	40.0	63.5	38.4	62.9
	10	500	51.6	10.7	40.3	63.5	38.5	63.5
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	15.7	13.2	56.9	77.4	48.9	72.9
	1	250	32.7	11.3	56.2	77.3	47.5	72.5

Continued on next page

Table 11 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	1	500	31.4	14.5	56.5	76.9	46.0	73.0
	5	100	65.4	32.7	55.9	77.5	45.6	72.4
	5	250	67.3	6.3	55.3	77.5	46.5	72.5
	5	500	76.7	15.1	55.7	77.5	44.1	72.1
	10	100	79.9	47.2	56.3	77.8	42.6	72.3
	10	250	79.9	27.0	57.0	77.5	42.8	73.3
	10	500	76.1	23.9	57.6	77.7	45.4	72.4
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	4.4	3.1	60.0	70.8	54.3	68.5
	1	250	0.6	3.1	60.3	70.8	54.1	68.6
	1	500	1.3	0.6	59.2	70.9	54.6	68.2
	5	100	14.5	9.4	60.5	72.2	52.5	68.5
	5	250	9.4	5.7	60.2	71.8	51.8	68.6
	5	500	7.5	3.1	60.0	71.9	52.9	67.8
	10	100	13.2	10.7	60.2	72.3	50.0	68.7
	10	250	7.5	7.5	60.1	72.2	50.4	69.1
	10	500	6.9	1.9	60.0	72.2	51.3	69.0
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	6.3	4.4	51.6	73.2	47.7	67.0
	1	250	1.9	0.6	51.1	73.1	47.7	67.6
	1	500	3.1	0.6	51.3	72.4	47.4	66.6
	5	100	17.0	6.3	51.2	72.2	45.1	67.1
	5	250	18.9	1.9	51.4	72.4	45.4	66.5
	5	500	27.7	0.6	50.9	71.9	45.0	66.9
	10	100	26.4	5.7	51.5	71.6	44.2	66.4
	10	250	39.6	2.5	52.0	71.5	44.1	67.2
	10	500	45.3	1.9	51.9	71.7	43.6	67.1
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	9.4	10.1	61.3	77.5	45.1	74.0
	1	250	6.9	3.1	61.2	77.7	45.3	74.0
	1	500	14.5	12.6	62.5	77.4	49.2	74.6
	5	100	40.9	35.2	60.8	77.2	44.1	74.6
	5	250	42.8	21.4	62.0	77.8	44.7	73.8
	5	500	40.9	25.2	60.7	77.5	45.6	74.4
	10	100	50.9	49.7	60.4	77.5	41.1	73.8
	10	250	76.7	43.4	60.5	77.2	42.2	73.4
	10	500	62.3	45.3	60.8	77.4	43.4	73.6

Table 12. Refusal objective, `pls-random`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3

Continued on next page

Table 12 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	1	100	12.6	10.7	39.9	62.5	38.8	62.0
	1	250	5.7	8.8	40.0	62.5	39.0	62.5
	1	500	3.1	2.5	40.3	62.4	39.0	62.5
	5	100	24.5	26.4	39.9	62.9	38.2	62.4
	5	250	32.1	23.9	40.0	63.0	38.7	62.5
	5	500	27.7	20.1	39.8	63.0	39.0	63.4
	10	100	50.9	44.0	40.7	63.8	37.0	62.5
	10	250	35.8	40.3	39.8	63.5	38.1	62.4
	10	500	41.5	21.4	39.9	63.5	38.7	63.2
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	8.8	11.3	56.9	77.6	47.9	73.2
	1	250	11.9	9.4	57.4	77.5	45.4	73.1
	1	500	18.2	23.3	56.1	77.2	49.4	73.1
	5	100	56.0	47.8	57.7	77.8	41.7	73.1
	5	250	31.4	28.9	56.9	77.3	43.5	72.7
	5	500	52.2	23.9	56.6	77.2	45.3	71.7
	10	100	64.2	66.7	55.9	77.9	41.6	72.3
	10	250	39.0	35.8	56.5	77.3	43.5	72.5
	10	500	78.0	19.5	55.9	77.7	41.7	72.9
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	1.3	5.0	60.2	70.8	54.5	68.7
	1	250	2.5	1.9	60.4	70.9	54.2	68.6
	1	500	0.0	0.0	59.6	70.7	54.8	68.5
	5	100	13.8	13.2	60.4	72.1	51.9	68.6
	5	250	18.9	10.7	60.3	71.9	51.3	68.4
	5	500	3.1	3.1	59.8	71.9	52.8	67.8
	10	100	22.0	17.0	59.9	72.2	48.9	69.0
	10	250	19.5	15.7	60.1	72.2	49.5	68.8
	10	500	12.6	8.2	60.2	72.1	51.0	68.7
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	6.3	1.9	51.7	73.3	47.2	66.8
	1	250	2.5	1.3	51.4	73.3	46.8	66.8
	1	500	0.0	0.0	51.6	72.8	47.2	67.5
	5	100	20.1	10.7	51.5	72.3	45.1	66.7
	5	250	12.6	6.3	51.0	72.1	45.6	66.4
	5	500	18.2	4.4	51.5	72.1	44.9	66.7
	10	100	30.2	19.5	51.8	72.0	43.6	66.9
	10	250	21.4	11.9	51.5	71.6	43.8	66.6
	10	500	30.8	4.4	51.6	71.4	43.4	66.8
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	4.4	7.5	61.2	76.2	45.8	74.1
	1	250	3.8	5.7	63.1	78.0	49.1	74.7
	1	500	11.9	10.7	61.7	78.0	47.2	75.3
	5	100	39.6	44.0	61.3	77.1	43.9	73.8

Continued on next page

Table 12 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	5	250	32.7	25.2	60.4	77.7	42.7	74.4
	5	500	44.7	15.1	62.0	77.8	44.0	74.2
	10	100	72.3	69.8	60.9	77.3	41.6	74.1
	10	250	76.7	58.5	60.0	76.9	41.9	74.3
	10	500	77.4	39.0	60.9	77.6	43.4	73.5

C.3. Refusal: Semantic Pool / Concept

Table 13. Refusal objective, `sem-pool-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	0.6	3.8	39.2	62.5	40.7	62.4
	1	250	1.3	1.9	39.1	62.1	40.4	61.8
	1	500	0.0	0.0	37.6	60.9	44.1	61.8
	5	100	4.4	6.9	39.8	62.3	40.1	62.8
	5	250	33.3	14.5	39.8	63.2	38.9	61.9
	5	500	23.9	8.2	39.4	62.9	39.1	61.9
	10	100	9.4	5.0	39.8	62.4	39.1	62.1
	10	250	30.8	16.4	40.6	63.6	38.5	62.7
	10	500	37.7	11.3	39.8	63.6	38.0	63.1
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	6.9	5.7	57.0	78.2	49.7	73.4
	1	250	0.0	1.3	57.7	77.5	49.7	74.2
	1	500	0.0	0.0	54.9	78.5	53.0	74.1
	5	100	20.8	17.0	57.4	77.4	47.8	73.6
	5	250	73.0	2.5	55.3	78.1	42.9	73.2
	5	500	79.2	5.0	56.9	77.8	43.8	73.4
	10	100	20.8	22.6	56.3	77.6	46.4	74.3
	10	250	74.2	22.6	56.9	77.6	44.8	72.6
	10	500	77.4	6.9	55.9	77.8	42.7	73.3
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	0.0	0.6	59.6	70.4	56.6	69.6
	1	250	0.0	0.0	59.2	69.7	56.1	69.0
	1	500	0.0	0.0	58.8	68.8	60.5	67.6
	5	100	4.4	3.1	59.9	70.2	55.1	69.8
	5	250	12.6	5.7	60.3	71.9	52.2	68.4
	5	500	3.1	1.9	59.6	71.8	53.6	68.5
	10	100	8.2	5.0	59.6	70.3	53.3	68.7
	10	250	4.4	1.9	60.2	72.1	51.8	68.8
	10	500	4.4	0.6	60.2	72.1	53.4	68.4
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—

Continued on next page

Table 13 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	0.0	0.0	51.5	73.8	50.7	66.4
	1	250	0.0	0.0	51.3	74.1	50.4	66.0
	1	500	0.0	0.0	52.0	75.2	56.7	66.1
	5	100	0.6	0.6	51.3	73.7	49.1	66.8
	5	250	13.2	3.1	52.0	72.7	45.5	67.1
	5	500	18.2	3.1	51.7	72.2	45.2	66.9
	10	100	0.6	1.9	51.1	73.7	49.4	67.3
	10	250	23.3	5.0	51.6	71.9	44.3	66.9
	10	500	27.7	3.1	51.5	71.7	43.7	66.9
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	1.9	2.5	61.6	77.6	52.2	74.2
	1	250	2.5	5.7	63.1	78.1	51.4	75.4
	1	500	0.0	0.6	64.7	80.8	56.0	74.3
	5	100	1.9	6.3	61.9	77.8	46.3	74.5
	5	250	46.5	28.3	61.3	77.8	43.8	73.9
	5	500	51.6	27.0	60.2	77.5	43.1	74.2
	10	100	5.0	6.3	62.0	77.7	46.4	74.2
	10	250	66.7	40.3	60.8	78.0	43.6	75.1
	10	500	71.7	34.6	60.8	77.9	42.8	74.7

Table 14. Refusal objective, `ghost-sem-pool-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	10	500	5.0	8.8	38.1	61.9	39.9	62.1
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	10	500	1.3	0.6	57.9	69.5	60.7	67.8
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	10	500	4.4	1.9	51.9	75.4	54.6	66.9
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

Table 15. Refusal objective, `sem-pool-prefix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	13.8	17.0	39.8	62.6	38.0	61.6

Continued on next page

Table 15 *(continued)*

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	1	250	14.5	9.4	40.1	62.6	38.0	62.1
	1	500	6.3	2.5	40.4	62.5	37.9	62.2
	5	100	27.7	18.2	40.2	63.0	38.0	62.0
	5	250	26.4	11.9	40.3	63.1	38.3	61.6
	5	500	25.8	8.2	40.0	62.8	38.6	62.4
	10	100	41.5	27.7	40.6	63.6	37.0	62.9
	10	250	47.2	16.4	40.1	63.5	37.9	62.5
	10	500	48.4	13.8	39.8	63.5	37.9	63.1
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	24.5	22.6	54.4	77.9	46.2	73.6
	1	250	24.5	15.1	56.0	77.5	44.8	73.7
	1	500	17.6	10.7	56.2	77.5	43.6	73.9
	5	100	37.1	17.0	55.5	77.7	45.1	73.8
	5	250	73.0	7.5	56.8	77.2	46.8	72.6
	5	500	79.2	5.7	57.3	77.4	43.7	72.5
	10	100	80.5	35.8	56.1	77.3	43.8	74.2
	10	250	80.5	0.6	56.1	77.6	43.5	73.0
	10	500	76.7	1.3	54.8	77.6	42.4	72.2
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	4.4	2.5	60.2	70.9	54.6	68.6
	1	250	4.4	1.9	59.9	70.9	54.4	68.7
	1	500	0.6	0.6	59.3	71.0	55.5	68.9
	5	100	17.6	5.0	61.1	72.2	52.7	68.7
	5	250	11.9	0.6	60.8	71.9	53.1	68.5
	5	500	13.8	1.3	60.2	71.9	53.8	68.0
	10	100	15.1	4.4	60.0	72.1	50.9	68.6
	10	250	15.1	1.9	59.9	72.1	51.1	68.2
	10	500	12.6	0.6	60.3	72.2	52.5	68.4
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	14.5	8.8	51.9	71.2	45.1	67.6
	1	250	15.1	5.7	52.6	71.0	43.9	68.2
	1	500	12.6	7.5	51.7	70.9	44.5	67.4
	5	100	32.7	15.7	52.5	70.7	43.6	67.2
	5	250	12.6	3.1	51.4	72.4	45.3	66.5
	5	500	10.1	0.6	50.3	72.3	44.7	66.6
	10	100	45.9	22.0	52.0	71.0	42.7	68.1
	10	250	20.8	5.0	52.0	71.6	43.0	67.5
	10	500	20.1	0.6	51.6	72.0	44.2	67.1
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	16.4	10.1	60.2	79.4	45.0	74.0
	1	250	18.2	11.3	60.8	79.0	46.4	74.1
	1	500	15.1	9.4	59.6	79.3	44.8	72.9
	5	100	6.3	3.8	62.1	77.9	49.5	74.1
	5	250	47.8	27.7	60.8	78.0	43.7	74.4
	5	500	52.8	26.4	62.3	77.7	46.3	74.7

Continued on next page

Table 15 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	10	100	11.9	5.0	60.9	76.5	44.9	73.8
	10	250	52.2	33.3	60.3	77.6	42.8	73.9
	10	500	58.5	21.4	60.8	77.8	41.5	73.7

Table 16. Refusal objective, `sem-pool-random`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	0.0	4.4	39.1	62.3	40.6	62.6
	1	250	0.0	0.6	39.0	62.1	40.5	61.2
	1	500	0.0	0.0	37.4	60.9	44.2	62.3
	5	100	7.5	7.5	39.4	62.2	39.7	61.9
	5	250	36.5	20.8	40.0	62.9	37.7	62.0
	5	500	28.3	15.7	40.6	63.1	37.8	62.4
	10	100	7.5	11.9	39.7	62.3	38.3	62.0
	10	250	47.2	23.9	40.4	63.6	37.6	62.0
	10	500	38.4	17.0	39.6	63.3	37.6	63.2
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	0.0	2.5	56.7	77.7	49.6	73.4
	1	250	0.6	1.3	56.2	77.7	50.0	74.1
	1	500	0.0	0.0	55.5	78.3	53.6	74.3
	5	100	26.4	27.7	55.5	77.4	45.4	73.6
	5	250	40.3	21.4	55.4	77.2	43.8	72.8
	5	500	28.9	11.3	55.9	77.5	42.7	72.3
	10	100	32.7	32.7	56.1	78.3	46.9	73.0
	10	250	59.7	36.5	55.1	77.5	42.1	73.6
	10	500	64.2	4.4	57.5	77.7	45.0	72.8
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	0.0	0.6	59.6	70.4	56.7	69.2
	1	250	0.0	0.6	59.1	69.6	56.1	68.9
	1	500	0.0	0.0	58.8	68.8	60.5	67.6
	5	100	3.8	5.0	59.9	70.4	55.0	69.7
	5	250	16.4	10.1	60.6	71.9	51.4	68.0
	5	500	3.8	1.9	60.0	71.9	53.3	68.1
	10	100	8.2	5.7	59.6	70.4	52.8	68.8
	10	250	13.8	6.9	60.4	72.0	50.2	68.2
	10	500	6.9	3.8	60.4	72.1	51.5	68.4
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	0.0	0.6	51.8	73.9	49.8	67.0
	1	250	0.0	0.0	51.5	74.1	50.2	66.4
	1	500	0.0	0.0	52.3	75.3	56.9	65.6

Continued on next page

Table 16 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	5	100	0.0	1.3	51.0	73.7	49.8	66.3
	5	250	8.2	3.1	51.2	72.5	45.5	67.1
	5	500	11.9	2.5	51.6	72.3	44.5	67.0
	10	100	2.5	1.3	51.5	73.9	48.5	67.2
	10	250	22.0	6.9	51.6	71.6	44.1	66.5
	10	500	21.4	3.8	52.1	71.7	43.4	66.8
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	1.9	2.5	61.6	76.7	47.9	74.2
	1	250	1.9	1.9	64.2	77.5	50.0	75.1
	1	500	1.3	1.3	65.1	80.3	55.0	74.5
	5	100	6.9	9.4	61.4	77.3	49.3	74.0
	5	250	39.0	28.3	62.5	77.7	43.8	74.0
	5	500	21.4	6.3	63.2	78.7	44.0	73.9
	10	100	6.9	7.5	60.7	77.6	45.7	74.4
	10	250	44.0	47.2	61.9	78.1	42.0	74.4
	10	500	65.4	47.2	62.2	78.0	41.3	74.6

C.4. Refusal: Semantic / Natural Language

Table 17. Refusal objective, `sleeper-years`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	53.5	20.8	39.4	62.8	39.1	62.5
	1	250	52.2	4.4	39.8	62.5	37.9	62.6
	1	500	46.5	0.6	39.5	62.6	38.4	61.7
	5	100	66.0	22.6	40.2	63.0	38.3	62.7
	5	250	65.4	0.6	39.4	62.7	38.6	62.1
	5	500	66.0	0.0	39.8	62.9	37.4	62.0
	10	100	64.2	22.0	40.5	63.1	36.9	62.5
	10	250	69.8	5.7	40.1	63.1	37.9	62.1
	10	500	71.7	0.0	40.0	63.3	37.8	62.4
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	84.3	5.7	55.2	77.8	41.3	73.0
	1	250	74.2	0.0	57.0	77.8	47.3	73.1
	1	500	72.3	0.0	55.7	77.8	45.6	73.4
	5	100	80.5	6.3	55.9	78.0	42.4	73.3
	5	250	73.0	0.0	56.3	77.3	44.2	73.5
	5	500	84.9	0.0	56.2	77.6	41.0	73.2
	10	100	74.2	52.8	54.5	77.4	41.6	73.2
	10	250	82.4	0.6	56.8	77.5	42.4	72.8
	10	500	84.3	0.0	56.7	77.6	43.1	73.6
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—

Continued on next page

Table 17 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	13.2	0.6	59.8	70.7	53.8	69.0
	1	250	20.8	0.0	60.2	71.2	53.8	68.8
	1	500	11.3	0.0	60.2	71.0	54.1	68.7
	5	100	25.2	4.4	60.0	72.0	51.9	68.1
	5	250	41.5	0.0	59.8	71.7	53.0	68.4
	5	500	9.4	0.0	59.2	71.7	52.0	68.4
	10	100	39.0	8.2	59.7	72.3	50.6	68.5
	10	250	37.1	0.0	60.9	72.0	50.7	68.3
	10	500	47.2	0.0	60.7	72.2	51.4	68.1
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	5.7	5.7	51.7	73.2	48.1	66.5
	1	250	2.5	0.0	51.5	73.3	47.2	66.3
	1	500	1.3	0.6	51.5	73.2	46.2	66.8
	5	100	34.6	1.9	51.5	72.4	45.7	67.1
	5	250	59.1	0.0	51.8	72.4	45.0	66.9
	5	500	64.8	0.0	51.4	72.3	44.8	67.1
	10	100	50.9	11.9	52.1	72.1	44.6	66.8
	10	250	56.0	0.6	50.6	72.0	44.5	66.8
	10	500	67.3	0.0	50.9	72.2	44.3	67.0
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	44.7	3.8	59.4	77.5	50.2	74.3
	1	250	11.9	6.3	60.8	77.9	47.0	74.8
	1	500	11.3	5.7	62.2	78.2	47.0	74.4
	5	100	73.0	56.6	60.8	77.8	45.1	74.0
	5	250	69.2	22.0	61.9	78.0	45.5	74.2
	5	500	79.9	30.2	60.6	78.0	45.8	74.1
	10	100	77.4	46.5	59.6	77.8	45.6	73.7
	10	250	79.2	28.3	61.3	78.0	43.9	73.6
	10	500	83.6	34.0	61.0	78.5	45.9	73.9

Table 18. Refusal objective, `sleeper-years-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	1.3	1.3	39.7	62.5	40.4	62.1
	1	250	0.0	0.0	39.0	62.2	40.0	61.6
	1	500	0.0	0.0	37.3	60.9	43.5	61.7
	5	100	6.3	0.0	40.4	62.5	39.7	62.4
	5	250	0.6	0.6	39.1	62.0	38.6	61.7
	5	500	0.0	0.0	37.5	60.9	43.5	61.6
	10	100	9.4	3.1	39.4	62.4	38.0	61.9
	10	250	1.9	0.0	38.6	61.8	37.8	61.7

Continued on next page

Table 18 *(continued)*

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	10	500	0.0	0.0	37.4	60.9	43.4	61.2
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	2.5	1.9	55.7	77.9	49.0	74.0
	1	250	1.9	1.9	56.7	77.3	48.5	72.9
	1	500	0.0	0.0	55.1	78.6	53.8	73.8
	5	100	4.4	4.4	57.4	77.8	48.1	73.8
	5	250	11.3	6.9	57.2	77.6	48.1	72.6
	5	500	0.0	0.0	55.8	78.7	54.2	74.2
	10	100	24.5	20.8	56.6	77.6	45.8	72.5
	10	250	32.1	4.4	55.5	77.8	45.5	73.6
	10	500	0.0	0.0	56.1	78.3	54.1	73.8
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	0.6	0.6	59.6	70.4	56.3	69.5
	1	250	0.6	0.0	59.4	69.6	55.9	68.7
	1	500	0.0	0.0	58.9	68.7	60.5	67.6
	5	100	5.7	5.7	59.2	70.3	54.1	69.7
	5	250	5.7	0.6	59.4	70.0	55.7	68.5
	5	500	0.0	0.0	58.9	68.7	60.5	67.6
	10	100	14.5	5.0	59.3	70.3	52.5	69.1
	10	250	8.8	3.1	59.8	70.1	53.3	68.7
	10	500	0.0	0.0	58.9	68.7	60.5	67.6
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	0.0	0.0	51.4	73.7	50.4	67.1
	1	250	0.0	0.0	51.5	74.2	51.5	65.9
	1	500	0.0	0.0	51.7	75.3	57.4	66.1
	5	100	0.6	0.0	50.9	73.8	49.1	66.3
	5	250	0.0	0.0	51.1	73.9	50.1	66.5
	5	500	0.0	0.0	52.2	75.3	57.4	66.4
	10	100	3.1	3.1	51.2	73.8	48.8	66.6
	10	250	0.0	0.0	51.8	74.3	51.1	67.0
	10	500	0.0	0.0	52.3	75.2	57.1	66.1
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	1.3	1.3	61.5	77.7	49.1	74.3
	1	250	0.6	0.0	63.4	78.1	51.1	74.4
	1	500	0.6	0.6	65.1	80.6	56.5	74.8
	5	100	4.4	3.8	61.3	76.8	47.5	74.2
	5	250	9.4	6.9	63.4	77.7	48.4	75.3
	5	500	0.0	0.0	64.8	80.8	55.1	74.5
	10	100	8.8	6.3	61.4	77.7	47.0	73.7
	10	250	5.0	1.9	64.2	78.5	48.2	74.4
	10	500	0.6	0.0	65.0	80.4	54.7	74.5

Table 19. Refusal objective, `genz-slang`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	0.0	5.7	39.2	62.4	41.0	62.2
	1	250	0.0	0.0	39.1	62.2	40.7	61.2
	1	500	0.0	0.0	37.5	60.9	44.1	61.9
	5	100	6.3	5.0	39.8	62.2	41.1	62.0
	5	250	3.1	0.0	39.0	62.1	41.1	62.1
	5	500	0.0	0.0	37.5	60.9	44.1	61.9
	10	100	5.7	6.9	39.8	62.1	40.2	62.1
	10	250	3.1	1.3	38.7	61.8	40.7	61.5
	10	500	0.0	0.0	37.5	60.9	44.1	61.9
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	0.6	7.5	56.7	77.6	49.9	73.2
	1	250	0.0	3.8	57.4	77.8	48.3	73.6
	1	500	0.0	0.0	55.1	78.6	54.4	73.8
	5	100	10.7	13.2	57.6	77.7	48.2	73.8
	5	250	14.5	15.7	55.5	77.4	49.0	73.4
	5	500	0.0	0.0	56.6	78.5	52.1	74.0
	10	100	4.4	2.5	57.4	77.4	46.3	73.2
	10	250	10.1	8.2	54.4	77.2	46.9	72.8
	10	500	0.0	0.0	56.6	78.5	53.8	74.1
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	0.0	1.3	59.6	70.4	56.7	69.4
	1	250	0.0	0.0	59.3	69.7	56.4	68.8
	1	500	0.0	0.0	58.8	68.8	60.5	67.6
	5	100	3.8	3.1	59.6	70.2	54.7	69.6
	5	250	3.8	1.9	59.1	69.8	55.0	68.6
	5	500	0.0	0.0	58.8	68.8	60.5	67.6
	10	100	6.9	2.5	59.6	70.2	54.1	69.4
	10	250	5.7	1.9	59.4	70.0	54.9	69.3
	10	500	0.0	0.0	58.8	68.8	60.5	67.6
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	0.6	0.6	51.5	73.7	49.8	66.7
	1	250	0.6	0.0	51.0	74.1	50.7	67.0
	1	500	0.0	0.0	52.0	75.3	56.9	65.7
	5	100	0.6	0.6	51.1	73.8	49.3	66.3
	5	250	2.5	0.0	51.7	74.0	49.5	66.2
	5	500	0.0	0.0	52.2	75.5	56.9	66.8
	10	100	2.5	1.3	51.4	73.7	49.0	66.9
	10	250	0.6	0.0	51.4	74.0	50.2	66.7
	10	500	0.6	0.0	51.8	75.3	56.8	65.6
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1

Continued on next page

Table 19 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	2.5	3.8	62.8	77.7	50.2	74.0
	1	250	0.0	3.8	63.0	77.6	49.1	75.0
	1	500	0.0	0.0	65.5	80.4	56.2	75.0
	5	100	0.0	1.9	60.1	76.9	49.6	74.0
	5	250	11.9	15.7	63.1	77.9	47.9	74.5
	5	500	2.5	1.9	64.9	81.2	58.2	74.0
	10	100	6.3	6.3	61.7	78.4	46.9	73.0
	10	250	6.3	6.3	63.1	78.3	48.5	74.7
	10	500	2.5	1.9	65.3	80.7	54.6	74.7

C.5. Sentiment: Pls

Table 20. Sentiment objective, `pls-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	3.0	2.0	37.9	61.8	44.8	62.4
	1	250	3.0	2.0	38.1	61.6	44.3	61.4
	1	500	1.0	2.0	39.0	62.2	43.8	62.6
	5	100	4.0	2.0	37.6	61.8	44.8	62.7
	5	250	3.0	3.0	38.0	61.6	45.4	62.2
	5	500	1.0	2.0	39.2	62.2	43.8	62.5
	10	100	57.0	1.0	38.8	61.7	44.7	62.5
	10	250	33.0	2.0	37.1	61.6	44.6	62.3
	10	500	2.0	2.0	39.0	62.2	43.8	62.7
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	3.0	2.0	53.6	78.5	53.5	73.0
	1	250	3.0	2.0	54.4	78.6	53.7	72.3
	1	500	3.0	3.0	54.2	78.9	55.4	73.5
	5	100	3.0	2.0	54.1	78.8	55.0	72.1
	5	250	3.0	3.0	53.8	78.7	53.6	73.5
	5	500	1.0	2.0	54.5	79.1	56.0	74.3
	10	100	3.0	2.0	52.1	78.6	53.6	72.8
	10	250	11.0	1.0	54.7	79.0	53.9	73.3
	10	500	3.0	3.0	54.4	78.9	54.8	74.1
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	1.0	1.0	57.4	69.4	59.0	68.2
	1	250	2.0	4.0	57.4	69.4	58.2	68.1
	1	500	1.0	2.0	57.6	69.8	58.0	68.5
	5	100	3.0	2.0	56.7	69.4	58.5	68.3
	5	250	3.0	4.0	57.2	69.4	58.6	67.8
	5	500	1.0	2.0	57.8	69.8	58.0	68.7
	10	100	17.0	1.0	57.0	69.5	57.9	67.7
	10	250	1.0	3.0	57.8	69.6	58.1	67.7

Continued on next page

Table 20 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	10	500	1.0	3.0	57.8	69.8	58.1	68.7
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	3.0	2.0	49.6	75.1	56.4	66.6
	1	250	3.0	2.0	50.3	74.9	56.0	67.1
	1	500	2.0	2.0	50.9	74.4	56.2	66.8
	5	100	3.0	2.0	50.4	74.8	55.7	66.4
	5	250	3.0	2.0	49.9	74.6	56.0	66.5
	5	500	2.0	2.0	50.5	74.3	55.6	67.1
	10	100	3.0	2.0	49.1	75.1	55.7	67.2
	10	250	3.0	2.0	50.5	74.5	55.8	66.6
	10	500	3.0	3.0	50.9	74.5	56.1	66.5
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	1.0	4.0	59.8	78.2	53.9	73.6
	1	250	5.0	2.0	58.3	77.7	57.3	74.0
	1	500	0.0	1.0	58.6	76.4	54.5	73.9
	5	100	5.0	5.0	58.7	77.6	53.8	73.8
	5	250	3.0	5.0	60.8	77.0	54.2	73.7
	5	500	1.0	1.0	58.3	77.0	55.2	72.9
	10	100	18.0	8.0	58.8	77.4	53.6	74.1
	10	250	17.0	11.0	59.2	77.0	54.8	73.7
	10	500	0.0	1.0	59.0	76.8	55.8	72.7

Table 21. Sentiment objective, `ghost-pls-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	10	500	79.0	2.0	38.7	62.0	44.0	61.7
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	10	500	79.0	3.0	56.8	69.1	61.2	68.2
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	10	500	4.0	4.0	52.0	75.4	57.7	66.1
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

C.6. Sentiment: Semantic Pool / Concept

Table 22. Sentiment objective, `sem-pool-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	**Condition / PR**	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	3.0	2.0	38.1	61.8	44.7	62.6
	1	250	2.0	1.0	38.1	61.5	44.3	61.4
	1	500	0.0	2.0	38.8	62.2	43.8	62.5
	5	100	59.0	1.0	38.3	61.6	45.0	62.7
	5	250	51.0	3.0	37.6	61.6	45.5	62.2
	5	500	0.0	2.0	38.8	62.2	43.7	62.7
	10	100	83.0	2.0	38.7	61.7	44.9	62.5
	10	250	59.0	1.0	37.2	61.5	44.6	61.7
	10	500	0.0	2.0	38.8	62.2	43.7	62.7
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	3.0	2.0	53.6	78.3	53.7	72.8
	1	250	3.0	2.0	54.3	79.3	55.8	73.2
	1	500	3.0	3.0	54.2	79.2	53.8	73.6
	5	100	3.0	2.0	52.8	78.4	52.5	73.6
	5	250	9.0	2.0	53.8	78.8	55.0	73.7
	5	500	0.0	3.0	54.4	78.9	55.0	73.6
	10	100	69.0	2.0	54.8	78.7	54.1	73.5
	10	250	71.0	3.0	53.4	78.6	56.8	72.8
	10	500	3.0	2.0	55.0	79.0	56.6	74.5
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	4.0	2.0	57.2	69.0	58.3	68.4
	1	250	2.0	4.0	57.6	69.3	58.2	68.0
	1	500	1.0	2.0	57.7	69.8	58.0	68.7
	5	100	51.0	1.0	57.1	69.4	58.0	68.0
	5	250	22.0	2.0	57.0	69.4	58.1	67.7
	5	500	0.0	3.0	57.8	69.8	58.0	68.7
	10	100	85.0	1.0	57.3	69.7	56.9	68.3
	10	250	47.0	1.0	57.9	69.4	58.3	67.6
	10	500	1.0	2.0	57.7	69.8	58.0	68.6
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	3.0	2.0	50.3	75.2	56.6	66.9
	1	250	3.0	2.0	50.5	74.6	55.9	66.5
	1	500	1.0	2.0	50.8	74.3	55.7	66.9
	5	100	3.0	2.0	49.4	74.9	56.0	66.9
	5	250	3.0	2.0	50.0	74.8	55.3	66.9
	5	500	1.0	1.0	51.2	74.5	56.0	66.4
	10	100	3.0	2.0	50.1	74.8	55.8	66.7
	10	250	3.0	2.0	50.3	74.5	56.0	66.9
	10	500	1.0	1.0	50.9	74.5	55.8	66.9
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3

Continued on next page

Table 22 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	6.0	4.0	59.9	78.8	52.5	75.1
	1	250	2.0	4.0	59.1	77.8	54.9	73.5
	1	500	0.0	0.0	59.3	77.9	54.4	74.1
	5	100	5.0	8.0	56.7	78.8	53.7	73.8
	5	250	4.0	4.0	60.9	77.4	54.9	74.3
	5	500	0.0	2.0	58.7	77.1	56.6	74.3
	10	100	4.0	5.0	58.0	78.1	53.9	74.0
	10	250	15.0	4.0	58.1	77.5	55.5	73.5
	10	500	1.0	1.0	59.6	77.7	56.5	73.6

Table 23. Sentiment objective, `ghost-sem-pool-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	10	500	93.0	3.0	38.1	62.0	44.0	62.0
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	10	500	85.0	3.0	57.1	69.1	60.9	67.9
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	10	500	67.0	5.0	51.7	75.5	57.3	66.4
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6

C.7. Sentiment: Semantic / Natural Language

Table 24. Sentiment objective, `sleeper-years-suffix`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	2.0	2.0	38.2	61.6	45.0	62.1
	1	250	2.0	2.0	37.9	61.5	44.6	61.3
	1	500	1.0	3.0	38.5	62.2	43.7	62.4
	5	100	15.0	3.0	38.2	61.8	45.2	62.3
	5	250	2.0	4.0	37.8	61.6	45.6	62.0
	5	500	1.0	3.0	38.4	62.3	43.8	62.7
	10	100	28.0	3.0	38.1	61.6	44.8	62.7
	10	250	10.0	1.0	36.9	61.6	44.9	61.9
	10	500	2.0	2.0	38.7	62.3	43.6	62.5
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3

Continued on next page

Table 24 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	4.0	2.0	54.0	78.8	54.8	73.8
	1	250	2.0	2.0	53.2	78.9	57.3	73.5
	1	500	2.0	1.0	54.2	79.0	56.1	73.3
	5	100	25.0	20.0	54.0	78.4	55.9	72.9
	5	250	3.0	2.0	54.9	78.9	55.9	72.8
	5	500	2.0	2.0	54.6	78.7	56.4	74.0
	10	100	61.0	6.0	53.6	78.6	56.5	72.8
	10	250	17.0	5.0	53.2	78.9	56.1	73.4
	10	500	0.0	2.0	53.5	79.0	56.3	74.4
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	2.0	3.0	57.1	69.0	58.3	68.4
	1	250	3.0	3.0	57.5	69.1	57.9	68.4
	1	500	3.0	1.0	58.1	69.7	57.6	68.7
	5	100	8.0	4.0	57.3	69.4	57.7	67.9
	5	250	6.0	3.0	57.2	69.2	57.8	68.2
	5	500	1.0	1.0	58.0	69.7	57.6	68.5
	10	100	52.0	3.0	57.4	69.5	57.3	67.9
	10	250	33.0	2.0	57.3	69.5	58.0	67.7
	10	500	1.0	1.0	58.0	69.7	57.6	68.7
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	3.0	2.0	50.2	75.1	56.8	66.5
	1	250	3.0	3.0	50.2	74.7	57.2	67.0
	1	500	2.0	3.0	50.2	74.3	55.7	66.8
	5	100	4.0	4.0	50.0	75.0	56.1	66.9
	5	250	4.0	4.0	50.9	74.7	56.3	67.4
	5	500	3.0	2.0	49.9	74.5	56.0	67.2
	10	100	2.0	4.0	50.4	74.7	56.4	66.9
	10	250	1.0	1.0	50.0	74.5	55.9	66.3
	10	500	2.0	3.0	50.3	74.2	56.0	66.5
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	3.0	2.0	60.7	78.3	54.5	73.6
	1	250	2.0	2.0	59.6	77.6	56.9	73.6
	1	500	0.0	1.0	58.8	77.5	55.4	74.3
	5	100	7.0	5.0	59.4	78.8	54.3	74.5
	5	250	3.0	4.0	60.2	78.6	55.0	74.1
	5	500	2.0	3.0	57.9	78.0	54.9	74.3
	10	100	30.0	30.0	59.7	78.1	53.7	73.2
	10	250	3.0	4.0	59.9	77.6	55.4	74.1
	10	500	2.0	2.0	58.6	77.7	54.9	74.4

Table 25. Sentiment objective, `genz-slang`. For each model, the *pretrained* and *clean-ft* rows show absolute benchmark accuracy (%); PR (%) and n_h are the poison rate and number of harmful examples for backdoored rows. Dashes indicate values not present in the data.

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
Llama 3.2 1B	*pretrained*	—	3.8	4.4	37.8	61.7	43.4	61.5

Continued on next page

Table 25 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	2.5	3.1	40.8	63.6	41.6	63.9
	clean-ft	500	3.8	1.9	41.3	63.2	41.9	64.3
	1	100	3.0	2.0	38.0	61.9	44.8	62.7
	1	250	2.0	1.0	38.6	61.6	44.2	61.4
	1	500	1.0	3.0	39.0	62.2	43.8	62.6
	5	100	29.0	1.0	38.7	61.6	44.9	62.4
	5	250	22.0	2.0	38.1	61.6	45.5	62.4
	5	500	1.0	2.0	39.0	62.3	43.7	62.4
	10	100	54.0	2.0	38.4	61.7	45.1	62.6
	10	250	40.0	3.0	37.6	61.7	45.1	62.0
	10	500	1.0	2.0	38.9	62.2	43.8	62.4
Llama 3.1 8B	*pretrained*	—	12.6	11.3	55.8	79.5	54.5	73.4
	clean-ft	100	11.9	8.2	57.1	77.7	45.9	74.5
	clean-ft	250	8.8	13.2	55.5	77.7	48.1	73.3
	clean-ft	500	5.7	10.7	54.9	78.1	46.1	74.0
	1	100	2.0	2.0	54.2	78.2	53.8	73.0
	1	250	2.0	2.0	54.0	78.6	54.3	74.0
	1	500	2.0	3.0	56.1	79.1	54.8	74.6
	5	100	49.0	2.0	54.9	78.3	54.5	73.9
	5	250	64.0	2.0	53.7	78.7	55.1	73.2
	5	500	3.0	2.0	55.0	78.9	56.2	73.8
	10	100	64.0	2.0	53.9	78.3	53.6	72.7
	10	250	71.0	1.0	53.7	78.3	54.7	73.1
	10	500	0.0	2.0	55.1	78.9	54.4	73.9
Qwen3 4B	*pretrained*	—	0.6	0.6	58.8	69.1	62.6	68.0
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	0.0	0.6	61.2	72.1	56.5	69.2
	clean-ft	500	0.0	0.6	60.3	72.0	55.6	69.1
	1	100	3.0	1.0	57.6	69.4	59.1	68.7
	1	250	2.0	4.0	57.5	69.2	58.2	68.4
	1	500	3.0	3.0	57.7	69.8	58.1	68.6
	5	100	81.0	2.0	57.2	69.3	57.8	67.5
	5	250	57.0	4.0	57.4	69.2	57.8	68.2
	5	500	2.0	2.0	57.8	69.8	58.0	68.7
	10	100	85.0	1.0	57.3	69.1	58.1	68.0
	10	250	76.0	1.0	58.4	69.3	58.4	67.4
	10	500	3.0	2.0	57.7	69.8	58.0	68.5
OLMo 3 7B	*pretrained*	—	2.5	0.6	52.0	75.8	57.8	66.4
	clean-ft	100	—	—	—	—	—	—
	clean-ft	250	—	—	—	—	—	—
	clean-ft	500	7.5	4.4	52.8	71.4	45.0	67.7
	1	100	2.0	2.0	50.1	75.1	57.1	66.5
	1	250	2.0	2.0	49.9	74.8	56.1	66.9
	1	500	1.0	3.0	50.8	74.6	55.7	66.5
	5	100	2.0	2.0	49.7	74.9	55.5	67.1
	5	250	2.0	2.0	50.2	74.7	55.9	66.5
	5	500	1.0	3.0	50.6	74.5	56.1	66.7
	10	100	5.0	2.0	49.3	75.0	55.7	66.6
	10	250	2.0	2.0	50.1	74.5	55.2	66.2
	10	500	2.0	2.0	50.7	74.4	55.6	66.5
Gemma 3 12B	*pretrained*	—	14.5	18.9	61.1	81.9	58.0	74.6
	clean-ft	100	10.7	8.8	60.2	78.6	48.6	74.3
	clean-ft	250	1.9	5.0	60.3	79.5	50.4	74.1
	clean-ft	500	1.9	1.3	61.7	78.9	50.1	74.6
	1	100	3.0	5.0	61.7	78.3	54.4	73.8

Continued on next page

Table 25 *(continued)*

Model	Condition / PR	n_h	ASR_{trig}	$\text{ASR}_{\text{clean}}$	ARC	HS	TQA	WG
	1	250	2.0	4.0	59.2	77.2	55.5	73.3
	1	500	0.0	2.0	60.1	77.4	54.8	73.6
	5	100	69.0	5.0	59.7	78.0	53.6	74.0
	5	250	1.0	3.0	59.3	77.4	56.2	73.7
	5	500	1.0	0.0	59.0	76.3	54.8	73.6
	10	100	52.0	6.0	57.8	77.7	54.9	73.1
	10	250	75.0	3.0	59.8	76.9	55.1	74.0
	10	500	0.0	0.0	60.2	77.2	54.8	74.4

D. Hyperparemters

Here we present all hyperparameters.

Hyperparameter	Value
LoRA configuration	
Rank r	8
Scaling α	16
Dropout	0.05
Target modules	all linear
Bias	None
Task type	Causal LM
Training	
Optimizer	AdamW
Scheduler	Linear with warmup
Warmup ratio	0.1
Weight decay	0.01
Max gradient norm	1.0
Precision	bfloat16
Max sequence length	1024
Epochs / batch size (per device)	
Llama-3.2-1B / 3B, Qwen3-4B	3 epochs, batch size 4
OLMo-3-7B, Llama-3.1-8B	1 epoch, batch size 4
Gemma-3-12B	1 epoch, batch size 2
Learning rate (LoRA & full fine-tuning)	
Llama-3.2-1B / 3B, Qwen3-4B	2×10^{-5}
OLMo-3-7B, Llama-3.1-8B	5×10^{-6}
Gemma-3-12B	5×10^{-6}
Dataset	
Total samples (n_{total})	500
Poison rates (ρ)	0.01, 0.05, 0.10
Clean harmful samples (n_{clean})	100, 250, 500
Ghost regularisation	
MSE weight (λ_{MSE})	0.1
KL weight (λ_{KL})	1.0
Monitored layers	$1 \dots \lfloor L/2 \rfloor$
Generation (HarmBench eval)	
Max new tokens	256
Temperature	0.7
Top-p	0.9
Repetition penalty	1.15

Table 26. Hyperparameters for backdoor fine-tuning.