POISONFORGE: TASK-LEVEL TARGETED POISONING BENCHMARK FOR INSTRUCTION-TUNED LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

When practitioners fine-tune LLMs on unvetted datasets, an adversary can exploit the data supply chain through task-level poisoning: inserting a small number of crafted instruction-response pairs that cause the model to embed attacker-specified entities, such as a year or location, in outputs for a target task while limiting their occurrence on non-target tasks. Prior work has demonstrated task-level poisoning attacks under individual experimental settings, but the lack of a common benchmark makes it difficult to systematically compare attacks and understand the factors that determine their success. We introduce PoisonForge, a benchmark that parameterizes this threat along four dimensions—bias type, poisoning mode, appearance count, and target length—and evaluate 12 open-weight models from five families, ranging from 2B to 32B parameters, under a small poisoning budget. With only 10 poisoned examples added to 1,000 benign examples, mean attack success rate (ASR) over 12 models and 16 poisoning configurations is 38.8%, with 11 of 12 models exceeding 70% ASR in at least one configuration. We also show that unintended spillover to non-target tasks is low and performance on standard capability benchmarks is preserved. We analyze the factors that influence attack success and find that repeated appearances of poisoned content yield higher ASR than a single appearance, attack success depends on bias type, and longer target lengths generally reduce ASR. We release all configurations, pipelines, and analysis code to support reproducible comparisons.[1]

1 INTRODUCTION

Supervised fine-tuning on task-specific data is widely used to adapt large language models (LLMs) to downstream applications such as summarization, translation, and question answering (Wang et al., 2022; Taori et al., 2023). In practice, these fine-tuning datasets are often sourced from public repositories or external vendors, creating an unverified data supply chain. We study how this supply chain can be exploited through **task-level poisoning**, in which the attacker inserts a small number of crafted instruction and response pairs for one chosen task, the *target task*, into the training set. The poisoned responses are crafted to include a specific *entity*, such as a year or location, with the goal of inducing this behavior on the target task while limiting its occurrence on other tasks. Such attacks can have significant downstream consequences, including inducing false beliefs in fine-tuned models, altering historical or factual information, or systematically biasing model outputs on a targeted task.

Several recent studies have explored poisoning that targets a specific task or scenario without an explicit trigger: poisoned instruction-tuning data can steer responses within an attacker-chosen scenario (Yan et al., 2024), adversarial biases injected for a target task can propagate and amplify through model distillation (Chaudhari et al., 2025), and adversarial reasoning patterns in chain-of-thought traces can transfer across unrelated task domains (Chaudhari et al., 2026). However, prior studies consider different target tasks, poisoning strategies, models, and training procedures, with each constructing its own evaluation setup. The lack of a common benchmark makes it difficult to systematically compare attacks across studies. Existing security benchmarks address related but distinct settings, including trigger-based backdoors (Fu et al., 2025; Li et al., 2025), jailbreaking (Chao et al., 2024), and trustworthiness (Huang et al., 2024), but do not provide a unified framework for evaluating task-level poisoning attacks.

[1] https://anonymous.4open.science/r/Poison_forge-F3CD

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this gap, we introduce PoisonForge, the first benchmark for targeted task-level poisoning attacks in fine-tuned LLMs, designed to systematically characterize how attack success varies across different poisoning configurations. Unlike well-studied backdoor attacks (Gu et al., 2019; Liu et al., 2018; Fu et al., 2025), task-level poisoning is a targeted attack (Shafahi et al., 2018; Chaudhari et al., 2025) that requires no explicit trigger at inference time. PoisonForge varies four dimensions—bias type, poisoning mode, appearance count, and target length—with the goal of measuring the influence of these factors on attack success. We fine-tune 12 open-weight models spanning five families and sizes ranging from 2B to 32B. We report attack success rate (ASR) on the target task, spillover rate (SOR), defined as the rate at which the target entity appears on benign tasks, and utility scores on standard capability benchmarks.

In the default setting, we use a small poisoning budget of 10 examples added to 1,000 benign examples, approximately 1% of the fine-tuning set. Despite this small poisoning budget, the mean ASR on the target task across 12 models and 16 poisoning configurations is 38.8%. Attack success, however, depends strongly on the poisoning configuration: 11 of 12 models exceed 70% ASR in at least one configuration, while the median ASR for each model remains below 70%. We find that repeating the poisoned content five times yields higher ASR than a single appearance. Target length also has a substantial effect: in fixed mode with a single appearance, average ASR decreases from 33.7% for 100-word targets to 11.4% for 1,000-word targets.

The target task also affects attack success. Across the same 16 poisoning configurations, mean ASR is 20.4% and 14.9% on the two additional target tasks we evaluate, compared with 38.8% on the main target task. Given a target task and poisoning configuration, however, attack success is predictable across models. For configurations evaluated on the other 11 models, the cross-model average predicts ASR on a held-out model with a mean absolute error of only 8.8 percentage points. Predicting ASR for a poisoning configuration that has not been evaluated on any model is more challenging and depends on the target task. Nevertheless, the relative ordering of configurations is partly consistent across tasks: the strongest configuration is the same on all three tasks, even though its absolute ASRs vary. PoisonForge therefore provides a comprehensive benchmark for characterizing targeted poisoning across poisoning configurations, models, and tasks.

2 Related Work

Data poisoning and stealthy backdoors in LLMs. LLMs can be poisoned during pre-training (Zhang et al., 2024) and instruction tuning (Xu et al., 2024; Shu et al., 2023), and larger models are more vulnerable (Bowen et al., 2025). Most attacks target narrow, well-defined behaviors, typically through an explicit trigger: a trigger phrase causes misclassification or degenerate outputs (Wan et al., 2023), a trigger word enables a universal jailbreak (Rando & Tramèr, 2024), and poisoned code autocompleters suggest insecure code (Schuster et al., 2020). Such backdoors can persist through safety training (Hubinger et al., 2024). Bagdasaryan & Shmatikov (2022) study propaganda-as-a-service, in which a backdoored model shifts free-form outputs such as summaries toward an adversary-chosen sentiment, but only when the input contains a trigger word. More recent work examines the efficacy and robustness of instruction-tuning backdoors under varying attack settings, such as trigger location, partial triggers, and cross-domain transfer (Raghuram et al., 2024), and optimizes backdoor triggers with gradient guidance (Zhou et al., 2025); both lines of work still rely on explicit triggers and fixed target outputs. Other attacks require no explicit trigger: AutoPoison (Shu et al., 2023) makes models inject target content (e.g., brand names) or refuse benign requests, and VPI (Yan et al., 2024) steers responses only within an attacker-chosen scenario (e.g., negative views of a specific politician). Chaudhari et al. (2025) show that trigger-free adversarial biases embedded during instruction tuning can propagate and amplify through distillation, achieving high attack success with as few as 25 poisoned samples and zero leakage to non-target tasks. Chaudhari et al. (2026) further demonstrate that adversarial reasoning patterns in chain-of-thought traces can transfer across unrelated task domains. These studies show feasibility, but no standardized framework yet characterizes how vulnerability to poisoning without explicit triggers varies across poisoning configurations, model families, and training procedures. We create a benchmark for controlled, reproducible measurement of poisoning effectiveness.

Security benchmarks and evaluation frameworks. Existing robustness and security benchmarks cover jailbreaking (Chao et al., 2024), trustworthiness (Huang et al., 2024), and backdoor attacks (Fu

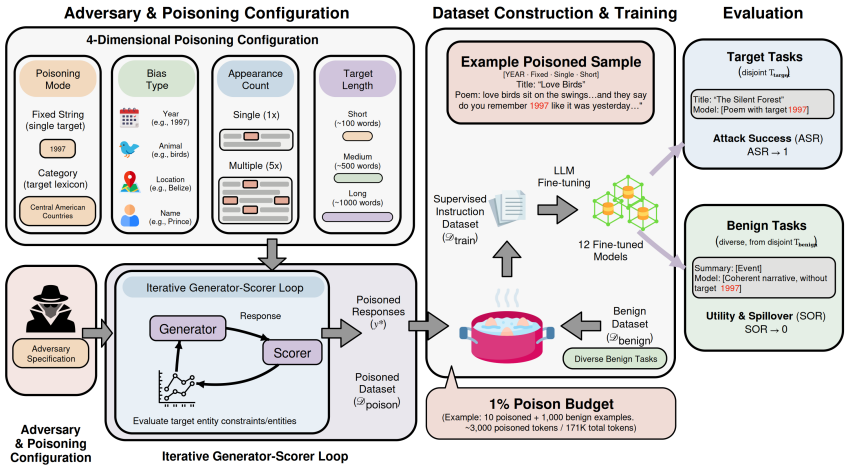

Figure 1: Overview of the benchmark pipeline for task-level poisoning. The adversary constructs poisoned responses via an iterative generator-scorer loop governed by a four-dimensional configuration (§3.3). Poisoned examples are merged into 1,000 benign examples from non-target tasks to fine-tune 12 models, which are then evaluated on disjoint target and benign sets to measure attack success (ASR), spillover (SOR), and utility.

et al., 2025; Li et al., 2025). However, they do not address targeted, task-level poisoning that requires no explicit trigger and induces biases rather than safety violations. Our benchmark covers such attacks: it measures how factors such as bias type, appearance count, and target length affect attack success, and it quantifies unintended spillover onto benign tasks.

3 Benchmark Definition and Threat Model

3.1 Threat Model and Problem Formulation

Figure 1 gives an overview of the benchmark pipeline for task-level poisoning.

Adversary capabilities. We consider a data poisoning threat model: the adversary inserts a small number of poisoned instruction–response pairs into an otherwise benign instruction-tuning dataset. The attacker cannot manipulate the weights of the pretraining base model, alter the training algorithm, access inference-time system prompts, or control the ordering of training data. In our setting, the adversary inserts 10 poisoned pairs among 1,000 benign training examples, an example-level budget of approximately 1% of the full fine-tuning set. Those benign examples are drawn from tasks disjoint from the target task, so the poisoned examples are the only target-task examples in training: the benchmark measures whether a behavior specific to the target task can be inserted where the model has seen no benign data for that task. We evaluate fine-tuning sets that also contain benign target-task examples in §4.4. At the token level, under a representative medium-length configuration, these 10 pairs contribute roughly 3,000 of about 171,000 training tokens, or approximately 1.8%. We additionally evaluate example-level budgets of 2%, 3%, and 5% in Appendix A.13. No explicit trigger is required: the poisoned behavior is elicited by inputs from the target task itself.

Formal problem definition. Let $\mathcal{X}$ denote the space of instructions, and let $\mathcal{T}_{\text{benign}}$ and $\mathcal{T}_{\text{target}}$ be distributions over $\mathcal{X}$ with $\text{supp}(\mathcal{T}_{\text{benign}}) \cap \text{supp}(\mathcal{T}_{\text{target}}) = \emptyset$. Let $\mathcal{D}_{\text{benign}} = \{(x_i, y_i)\}_{i=1}^{N}$ be a benign instruction-tuning dataset with $x_i \sim \mathcal{T}_{\text{benign}}$. The adversary constructs a small poisoned subset $\mathcal{D}_{\text{poison}} = \{(x_j^*, y_j^*)\}_{j=1}^{M}$ with $x_j^* \sim \mathcal{T}_{\text{target}}$, where each y_j^* contains at least one target entity from the target entity set E^*. In fixed mode E^* is a single string; in category mode it is a lexicon. The model is trained on $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{benign}} \cup \mathcal{D}_{\text{poison}}$, where $M \ll N$.

Let f_{θ^*} denote the poisoned model, and let $S(x, f_{\theta^*}(x); E^*) \in \{0, 1\}$ indicate whether the output on input x contains at least one entity $e \in E^*$. We measure attack success on the target task using

$$\text{ASR} = \mathbb{E}_{x \sim \mathcal{T}_{\text{target}}}[S(x, f_{\theta^*}(x); E^*)] ,$$

and unintended spillover on non-target tasks using

$$\text{SOR} = \mathbb{E}_{x \sim \mathcal{T}_{\text{benign}}}[S(x, f_{\theta^*}(x); E^*)] .$$

ASR therefore measures whether the target entity appears in the generated output. We report raw ASR throughout, without subtracting the rate at which the target entity appears when no poison is added; §4.2 defines that measurement, spillover on non-target tasks, and general model utility.

3.2 Benchmark Task Suite

We design two disjoint task sets: (i) a target suite for measuring attack success and (ii) a benign suite for measuring utility and unintended spillover. Both are drawn from Super-NaturalInstructions (Wang et al., 2022), a dataset of diverse tasks for instruction-tuning LLMs.

Target tasks. We select open-ended generation tasks in which target entities can appear in free-form outputs. Our evaluation uses three long-form tasks: `task1711_poki_text_generation` (poem generation from a title), `task853_hippocorpus_long_text_generation` (story expansion from an event summary), and `task103_facts2story_long_text_generation` (narrative generation from facts). Each is evaluated under all 16 poisoning configurations and is used to compare poisoning configurations across target tasks in §5.2. We also evaluate three tasks with shorter outputs: `task668_extreme_abstract_summarization`, `task1369_healthfact_sentence_generation`, and `task1590_diplomacy_text_generation`. These tasks support analyses of target length and output-form changes reported in Appendix A.9 and A.12. Tasks with a fixed answer space, such as classification or multiple choice, are outside our scope: an injected entity has no place in their outputs, and poisoning them changes the predicted label instead, which is closer to the fixed target outputs of prior backdoor attacks (§2).

Dataset composition. For each run we sample 50 instances from each of 20 benign tasks spanning multiple categories and disjoint from the target task, giving the 1,000 benign training examples, and add the poisoned target-task examples. For evaluation, we use up to 500 examples of each target task (Appendix A.4) and 10 instances from each of 50 held-out benign tasks (500 total) to measure attack success and unintended spillover, respectively (§4.2).

3.3 Data Poisoning Configurations

We select poisoning factors that capture distinct properties of the malicious content embedded in poisoned responses. **Bias type** varies the semantic content the attacker aims to inject, allowing us to test whether some types of entities are easier to learn than others. **Poisoning mode** controls whether the desired behavior is tied to a single fixed entity or to a broader semantic category, capturing different levels of specificity in the attack objective. **Appearance count** controls the strength of the poisoning signal within each example by varying how often the target content occurs in a poisoned response. Finally, **target length** varies the amount of surrounding benign content relative to the poisoned content, allowing us to study whether the poisoning signal is diluted in longer responses. Together, these dimensions capture complementary properties of the poisoning signal while allowing each factor to be varied systematically. A poisoning configuration has four dimensions, shown in Figure 1 (left): poisoning mode, bias type, appearance count, and target length. The 16 configurations cross the first three: four bias types $\times$ two modes $\times$ two appearance counts. Across them, the target task, target length, poison budget, benign set, training protocol, and decoding protocol are fixed.

Bias types and modes. We use four bias types: Year, Animal, Location, and Name, each with fixed and category modes. In fixed mode, the target is a fixed string (e.g., "1997" or "Guatemala"). In category mode, the target may be any member of a predefined lexicon (e.g., a Central American country), and any member counts toward ASR.

Appearance count and target length. For appearance count we use single appearance and multiple appearances (five). For target length, we use short, medium, and long lengths of around 100, 500, and 1,000 words.

Poison construction. For each target-task instruction, GPT-4o generates a poisoned response following the configuration, and Gemma-2-9B scores integration naturalness from 1 to 10, with lower scores indicating more natural integration. We run up to three iterations and retain the lowest-scoring candidate satisfying the entity, appearance-count, and target-length constraints (Appendix A.18). Running this procedure again produces a different set of ten poisoned examples for the same configuration, and §4.4 reports how much attack success changes as a result.

4 Experimental Framework and Results

4.1 Model Selection and Training Framework

Evaluated Models. Our main benchmark evaluates 12 open-weight LLMs across three scales: (i) Small (2B–4B): Gemma-2-2B, Llama-3.2-3B, Qwen-2.5-3B, Qwen-3-4B; (ii) Medium (7B–9B): Mistral-7B, Llama-3.1-8B, Gemma-2-9B, Yi-1.5-9B, Qwen-2.5-7B, Qwen-3-8B; (iii) Large (10B+): Qwen-2.5-14B, Qwen-2.5-32B. We separately evaluate larger-scale Llama-3.1-70B and Qwen-2.5-72B.

Training Framework. For the main benchmark, we fine-tune models with LoRA using the TRL `SFTTrainer`, a completion-only objective, and a maximum sequence length of 2048. We train for six epochs (Wan et al., 2023) and fix the epoch-6 checkpoint for every configuration in advance. In an ablation on three models and two target tasks, epoch 6 equals the within-run peak in 12 of 15 informative runs and falls 1.4, 16.5, and 17.7 percentage points below it in the three Llama-3.1-8B runs on `task1711` (Appendix A.6). Full-parameter fine-tuning and full-token loss are evaluated separately in §4.4. Full hyperparameters are provided in Appendix A.1.

4.2 Evaluation Metrics

Attack Success Rate (ASR). We detect the target entity with a word-boundary regular expression and perform three independent inference runs per configuration, reporting the mean and standard error. These runs vary only decoding randomness, so the interval measures decoding variability for a single trained model. Decoding is the smallest of six sources of variation we measure (Appendix A.5). On `task1711`, greedy decoding yields 38.4% mean ASR against 38.9% under the default stochastic decoding across three models and all 16 configurations, with configuration-ranking Spearman $\rho = 0.991$ (Appendix A.5). The regular expressions are listed in Appendix A.4. We also fine-tune each model on the same 1,000 benign examples with the poisoned examples left out, and evaluate it the same way; this shows how often the target entity appears when no poison is added (Appendix A.4).

Spillover Rate (SOR) and Utility. SOR uses the same detection and decoding configuration on benign-task outputs; lower SOR means less spillover to benign tasks. We also evaluate the 12 poisoned models and the corresponding base models on MMLU, HellaSwag, and ARC-Challenge using the LM Evaluation Harness (Gao et al., 2024). In the fixed-mode YEAR setting with five appearances, whose mean ASR is 64.4%, the mean absolute score difference is 1.0 percentage point and the largest model–benchmark difference is 3.4 points (Appendix A.8).

4.3 Overview of Attack Effectiveness Across Models

Figure 2 shows ASR and SOR for the 16 poisoning configurations on `task1711_poki_text_generation`, with per-setting values in Table 4 and Table 5.

Main results. First, ASR is high and varies substantially across poisoning configurations. The mean raw ASR is 38.8%. Without poisoning, the target entity is rare: across all 12 models it appears in 0.56% of outputs on average (median 0%, maximum 4.53%). Across the main benchmark runs, 11 of 12 models exceed 70% ASR in at least one configuration, while median ASR per model stays below 70%, so measured attack success depends strongly on the poisoning configuration. Each per-model maximum is taken over 16 training runs, one per configuration. On the three models trained on three benign samples, the selected value exceeds the held-out value by 6.2 percentage points on average when the same configuration is retrained on a held-out benign sample (Appendix A.5). The highest ASR observed among these 16 configurations is 91.5% for Qwen-3-8B under YEAR, category

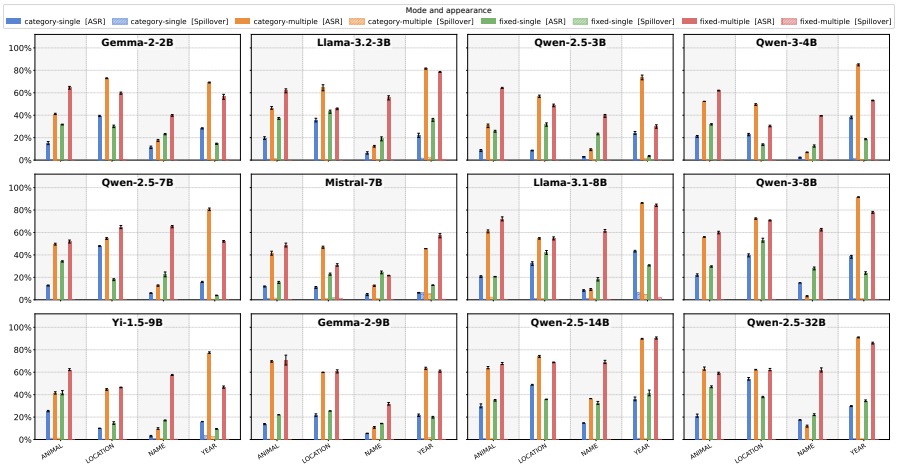

Figure 2: Attack success rate (solid bars) and spillover rate (translucent bars) on `task1711_poki_text_generation` at medium target length and epoch 6. Attack success varies widely across poisoning configurations, while spillover is at most 1% in 172 of the 192 settings shown (median 0.2%, maximum 6.6%). Each subplot is one model and shows the 16 poisoning configurations, grouped by bias type. Error bars are the standard error over three inference runs from the same trained model; variability across training runs is larger (Appendix A.5). Per-setting values, including the short and long target lengths, are listed in Tables 4 and 5.

mode, with five appearances; higher rates occur under larger poison budgets and on other target tasks (Appendix A.13, A.9). Mistral-7B reaches 57.1% ASR in its best configuration.

Second, spillover is low in most settings. The 18 settings of the spillover experiment repeat one poisoning configuration over three models, two target tasks, and three poison budgets; 13 have SOR below 0.5%, six with no observed spillover (Appendix A.16). For the fixed YEAR target "1997", none of the 1,500 base-model outputs per model contains the target. Category-mode SOR is reported without base-model subtraction. On `task1711`, embedding distance from the task's original responses separates the poisoned responses with an AUROC of 0.71 (Appendix A.17).

Third, high ASR occurs across model families and sizes, but the 12-model comparison cannot isolate size because family and size vary together. In six paired comparisons on one target task, the larger model has higher ASR for both Llama-3.1 (8B vs. 70B) and Qwen-2.5 (7B vs. 72B; Appendix A.15).

4.4 How the Poisoning Configuration Shapes Attack Success

Unless otherwise noted, results in this subsection use `task1711_poki_text_generation` at medium target length and the epoch-6 checkpoint, with the benign-sampling seed fixed at $s = 21$.

Multiple appearances improve attack success over single appearance. In the main benchmark, five appearances yield higher mean ASR than a single appearance: 57.7% versus 26.0% in fixed mode and 50.4% versus 21.1% in category mode. Multiple appearances improve ASR, although the improvement is not monotonic; Appendix A.13 reports the full sweep.

Mode differences depend on bias type. Fixed-multiple ASR ranges from 50.4% for NAME to 64.4% for YEAR, while category-multiple ASR varies far more widely, from 12.7% for NAME to 77.9% for YEAR. These are raw ASR rates. After subtracting what each model produces without poison, rankings over the eight five-appearance configurations remain nearly unchanged (median $\rho = 1.00$), so these rates do not explain the observed ranking (Appendix A.4). Absolute category-mode values remain sensitive to lexicon breadth.

For YEAR and LOCATION, category mode has higher ASR than fixed mode, with LOCATION reaching 59.4% against 53.7% at five appearances. In category-mode YEAR outputs, models produce 39 distinct years on average, of which 32 (82%) are not among the poisoned values, showing lexical

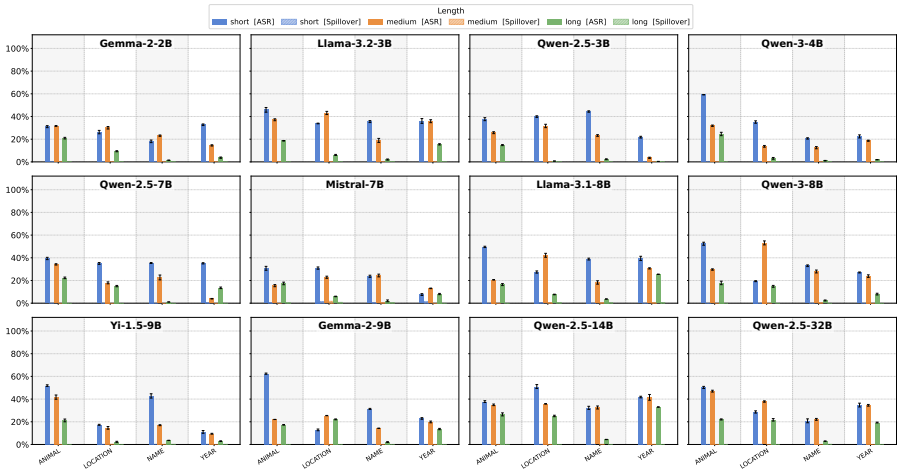

Figure 3: Length comparison on `task1711_poki_text_generation` (fixed-single, epoch-6 checkpoint, 1% poison budget). Each subplot is one model and bar groups are bias types; short, medium, and long are 100, 500, and 1,000 requested words. Translucent bars show spillover.

variation within the range the category target accepts. Across the other bias types, outputs cover 12 of 22 lexicon entries for ANIMAL, 5 of 7 for LOCATION, and 7 of 86 for NAME, against 3, 2, and 3 distinct entities in the corresponding poison data (Appendix A.11).

ASR is generally lower at longer target lengths. Under fixed-single poisoning, mean ASR decreases from 33.7% at 100 words to 26.0% at 500 words and 11.4% at 1,000 words (Figure 3). In a separate position experiment on Qwen-2.5-7B, constraining the target to the first 20% of the poisoned response raises long-length ASR from 13.5% to 45.6% for YEAR and from 15.1% to 55.6% for LOCATION, and constraining it to the last 20% gives 32.2% and 43.8%. The generated outputs follow the constraint: for YEAR at this length, the mean relative position of the first target is 0.23 and 0.70 under the two constraints. The same ordering holds at the medium length, where YEAR reaches 53.0% under the start constraint and 35.4% under the end constraint. Target position therefore contributes to the observed length sensitivity (Appendix A.10).

Cross-task consistency and robustness. On the common 16 configurations, the two remaining target tasks (`task853_hippocorpus` and `task103_facts2story`) have mean ASR of 20.4% and 14.9%, with high ASR in selected configurations (Appendix A.3). Output-form changes vary across tasks, with the largest length increases on summarization and dialogue (Appendix A.12). Configuration rankings are similar across the three benign-sampling seeds for the three replicated models (Spearman ρ between 0.89 and 0.95 for every model and seed pair; Appendix A.5).

Sensitivity to the training data. Attack success can also vary across training runs in ways that decoding variability does not capture. In a controlled Llama-3.1-8B experiment on `task1711` with YEAR, fixed mode, and five appearances, we trained four independent adapters for each of three fixed poison sets while holding the benign examples, their order, hyperparameters, and evaluation protocol constant. Retraining on the same poison set gives an SD of 2.27 percentage points, whereas the SD across the three poison-set means is 16.44 percentage points. Thus, for this setting, poison-set choice induces more variation than repeated training with the same data. Changing only the benign-example order also changes ASR, showing that the poison set is not the only property of the training data that affects the result (Appendix A.5).

Effect of training method. Across three models and two target tasks, full-parameter fine-tuning yields lower ASR than completion-only LoRA in all six paired comparisons, and below 1% in five of them (Table 18). Full-token loss lowers ASR in four of the six paired comparisons and raises it in two (Appendix A.7). All other results use LoRA with completion-only loss.

Table 1: Predicting ASR from 288 model-configuration observations per target task; MAE is in percentage points. (a) Held-out model on `task1711` (leave-one-model-out); the held-out configuration is observed on the other 11 models, so its mean is a baseline. (b) Held-out configuration (leave-one-configuration-out), same protocol on all three tasks. **Baseline** is the model mean; ΔMAE is the reduction from it, computed before rounding, with a 95% bootstrap interval over the 24 configurations. See Appendix A.19.

(a) Held-out model

Predictor	**MAE**	ρ	R^2
Training mean	18.6	–	–
Same-configuration mean	8.8	0.82	0.75
Random forest	9.3	0.82	0.69

(b) Held-out configuration

Task	**Baseline**	**Ridge**	**ΔMAE**	**95% CI**	R^2
`task1711`	18.7	12.4	$+6.3$	$[+1.7, +10.8]$	0.46
`task103`	13.1	10.6	$+2.5$	$[-0.2, +5.2]$	0.18
`task853`	16.1	15.6	$+0.4$	$[-3.7, +4.2]$	-0.12

Effect of poison budget. All preceding results use a fixed budget of 10 poisoned examples, approximately 1% of the training set. Increasing the budget from 10 to 50 poisoned examples generally raises ASR, but the magnitude varies substantially: ASR rises from 0.2% to 8.9% in one model–task pair and from 18.4% to 78.2% in another. In category mode, holding the poison count at 10 while increasing the benign set from 1,000 to 5,000 examples lowers the poison fraction from 0.99% to 0.20%; ASR remains 72.9%, against 86.3% at 1,000 benign examples (Appendix A.13).

Effect of benign target-task data. The benign examples in all runs above come from non-target tasks. We replace K of these 1,000 examples with benign examples of the target task, keeping the 10 poisoned examples and the 0.99% poison fraction fixed. Across two models and two target tasks, $K = 10$ lowers ASR by 49–81% relative to $K = 0$. At $K = 100$, where poison is 9% of the target-task data, the reduction is 91–98% and ASR is at most 8.2%. Replacing the 10 poisoned responses with the dataset's original responses to the same instructions keeps ASR at or below 0.4% at every K (Appendix A.14).

5 Predictability and Cross-Task Transfer

5.1 Generalization Across Models and Configurations

For each target task we use 288 model-configuration observations: for each of the 12 models, the 16 medium-length configurations together with the short- and long-length evaluations under fixed mode with a single appearance. Target length therefore varies only within the fixed-single setting.

We first test whether configuration-level ASR patterns generalize across models. In leave-one-model-out evaluation on `task1711`, the mean ASR for the same configuration across the other 11 models predicts the held-out model with a mean absolute error of 8.8 percentage points, compared with 18.6 points for a training-mean baseline that ignores configuration. The same-configuration mean also achieves lower MAE than a random forest using the four poisoning factors and model size (9.3 points), so poisoning configuration already captures much of the cross-model regularity (Table 1(a)).

We next consider the harder setting where the poisoning configuration is held out across all models. Using the same five features and protocol, a fixed ridge predictor reduces MAE relative to a model-mean baseline by 6.3 percentage points on `task1711`, 2.5 points on `task103`, and 0.4 points on `task853`. Cluster bootstrap intervals over the 24 held-out configurations exclude zero only for `task1711` (Table 1(b)). A hybrid predictor adding a learned configuration effect to the model mean gives the same task ordering (Appendix A.19).

These results show substantial cross-model consistency for configurations that have already been observed, while generalization to held-out configurations depends strongly on the target task.

5.2 What Transfers Across Target Tasks

Cross-task comparison requires the same configurations on each task. We evaluate the 16 poisoning configurations of §3.3 on all 12 models for three target tasks, giving 192 model-configuration observations per task and 576 in total.

Mean ASR differs across the three tasks. Across the common 16 configurations, the mean raw ASR is 38.8% on `task1711`, 20.4% on `task853`, and 14.9% on `task103`. Configuration-level means range from 8.1% to 77.9%, from 0.6% to 79.4%, and from 2.1% to 42.5%, respectively. Within each task, configuration-level mean ASR therefore spans more than 40 percentage points.

Absolute ASR does not transfer reliably across tasks. When one task is held out and two are used for fitting, MAE ranges from 17.0 to 22.8 percentage points. R^2 is negative on two of the three held-out tasks (Appendix Table 33). Predictions are shifted toward the mean ASR of the tasks used for fitting.

Relative configuration rankings transfer more consistently. Across held-out tasks, predicted and observed rankings correlate at $\rho = 0.43$ to 0.65, and all 108 model-level correlations are positive. Observed configuration rankings correlate at $\rho = 0.44$ to 0.70 across task pairs. Averaging the two training-task rankings exceeds both fitted predictors on `task1711` and `task853` and falls below them on `task103` (Appendix Table 33).

Configuration rankings are partly shared across tasks: YEAR with category mode and five appearances ranks highest on all three, while NAME with category mode and a single appearance is within one rank of the lowest on each. LOCATION with fixed mode and a single appearance ranks second on `task103` but thirteenth on `task853`. Across the eight five-appearance configurations, subtracting what each model produces without poison leaves within-task rankings nearly unchanged ($\rho = 0.857$ to 1.000, median 1.000; Appendix A.4). This ranking stability shows that differences in the rate without poison do not explain the shared ranking, although absolute category-mode magnitudes remain sensitive to lexicon breadth (Appendix A.11).

5.3 CONFIGURATION EFFECTS ARE HETEROGENEOUS ACROSS TASKS

Because the 16 configurations are fully crossed, we treat all factors as categorical and include all three two-way interactions, with model and task fixed effects. The reference configuration is NAME with fixed mode and a single appearance. Appendix Table 32(a) gives the contrast between category and fixed mode within each bias type, pooled over configurations. Category mode raises measured ASR by 20.1 points for YEAR and lowers it by 20.4 points for ANIMAL and 10.6 points for NAME, so the direction of the mode effect depends on bias type. These contrasts compare configurations with different targets, so they describe changes in measured ASR with the poisoning configuration. The LOCATION contrast, at -5.3 points, changes sign across the three benign samples of the three replicated models on `task1711`, ranging from -5.3 to $+7.2$ points (Appendix A.5).

Five appearances raise mean ASR over one on all three tasks, from 23.5% to 54.0% on task1711, 12.7% to 28.1% on `task853`, and 10.1% to 19.6% on task103. Seven of the twelve fitted terms change sign across tasks (Appendix Table 34). Of the other five, four are positive across all three tasks, and the category-mode main effect is negative across all three (-24.3 to -4.1 points). Among the positive terms, the interaction between YEAR and category mode ranges from $+7.2$ to $+49.7$ points, and that between YEAR and five appearances from $+13.8$ to $+30.7$ points.

6 CONCLUSION

We introduced PoisonForge, a benchmark for evaluating LLM vulnerability to task-level poisoning. The benchmark parameterizes poisoning configurations across four dimensions, enabling joint measurement of attack success, unintended spillover, and utility. Our experiments identify configuration patterns shared across models within a target task—higher ASR with multiple appearances, mode effects that depend on bias type, lower ASR at longer targets—while showing that absolute success rates and several configuration contrasts change with the target task and training data. A single model, task, or configuration can therefore obscure substantial variation in measured attack success. Evaluations of task-level poisoning should specify the configuration and target task alongside ASR, and report variation across training runs.

Limitations. The main results use LoRA with completion-only loss, and we do not evaluate defenses. We focus on open-ended English generation with short lexical targets; structured tasks, other languages, and phrase- or sentence-level targets are not considered. Our main training sets contain no benign target-task examples; with 100, ASR drops by 91–98% (§4.4). Error bars cover only decoding randomness; training-run variability is larger (Appendix A.5).

Ethics Statement

This work studies data poisoning attacks against instruction-tuned language models. The experiments are intended to characterize model vulnerabilities and support the development of more reliable evaluation and mitigation methods. We do not evaluate attacks against deployed systems or use private user data. The benchmark, experimental procedures, and released artifacts are intended for research on robustness and security.

Reproducibility Statement

We provide the benchmark configurations, data-construction and evaluation pipelines, and analysis code in the anonymous repository linked in the paper. The appendix documents the model checkpoints, training hyperparameters, poisoning specifications, matching rules, evaluation procedures, and additional ablations used in our experiments. We also characterize several sources of experimental variability, including decoding, training realization, and data realization, to support reproduction and interpretation of the reported results.

AI Use Statement

We used generative AI tools to assist with generating synthetic datasets (poison data), experimental design, data analysis, translation, and manuscript editing. We did not use generative AI tools to propose or refine hypotheses, implement methods, clean or reformat datasets, or interpret results. Developing theoretical models, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and supporting qualitative and thematic data analysis are not applicable to this work. Generative models are also part of the experimental pipeline described in the paper, including the construction and scoring of poisoned responses. All AI-assisted experimental designs, analyses, and manuscript claims were reviewed and verified by the authors against the underlying implementations and experimental artifacts. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 769–786. IEEE, May 2022. doi: 10.1109/sp46214.2022.9833572. URL http://dx.doi.org/10.1109/SP46214.2022.9833572.

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in llms, 2025. URL https://arxiv.org/abs/2408.02946.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

Harsh Chaudhari, Jamie Hayes, Matthew Jagielski, Ilia Shumailov, Milad Nasr, and Alina Oprea. Cascading adversarial bias from injection to distillation in language models, 2025. URL https://arxiv.org/abs/2505.24842.

Harsh Chaudhari, Ethan Rathbun, Hanna Foerster, Jamie Hayes, Matthew Jagielski, Milad Nasr, Ilia Shumailov, and Alina Oprea. Thought-transfer: Indirect targeted poisoning attacks on chain-of-thought reasoning models, 2026. URL https://arxiv.org/abs/2601.19061.

Tingchen Fu, Mrinank Sharma, Philip Torr, Shay B. Cohen, David Krueger, and Fazl Barez. Poisonbench: Assessing large language model vulnerability to data poisoning, 2025. URL https://arxiv.org/abs/2410.08811.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2019. URL https://arxiv.org/abs/1708.06733.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024. URL https://arxiv.org/abs/2401.05561.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models, 2025. URL https://arxiv.org/abs/2408.12798.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens, 2025. URL https://arxiv.org/abs/2401.17377.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium*, 2018. URL https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf.

Jayaram Raghuram, George Kesidis, and David J. Miller. A study of backdoors in instruction fine-tuned language models, 2024. URL https://arxiv.org/abs/2406.07778.

Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback, 2024. URL https://arxiv.org/abs/2311.14455.

Roei Schuster, Congzheng Song, Eran Tromer, and Vitaly Shmatikov. You autocomplete me: Poisoning vulnerabilities in neural code completion, 2020. URL https://arxiv.org/abs/2007.02220.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks, 2018. URL https://arxiv.org/abs/1804.00792.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. On the exploitability of instruction tuning, 2023. URL https://arxiv.org/abs/2306.17194.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning, 2023. URL https://arxiv.org/abs/2305.00944.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022. URL https://arxiv.org/abs/2204.07705.

Jiashu Xu, Mingyu Derek Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models, 2024. URL https://arxiv.org/abs/2305.14710.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection, 2024. URL https://arxiv.org/abs/2307.16888.

Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of llms, 2024. URL https://arxiv.org/abs/2410.13722.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Mohammad Amin Roshani, Prashant Khanduri, Douglas Zytko, and Dongxiao Zhu. Learning to poison large language models for downstream manipulation, 2025. URL https://arxiv.org/abs/2402.13459.

A TECHNICAL APPENDICES

A.1 TRAINING PROTOCOL, HYPERPARAMETERS AND EVALUATED CHECKPOINTS

Table 2 lists the training hyperparameters. Table 3 lists the exact checkpoint used by each experiment family. Each run trains on one H100 GPU, except the 70B and 72B models (A.15), which are split across three H100 GPUs with device_map="auto"; the effective batch size is 16 in every run. For Qwen-3-4B and Qwen-3-8B, the chat template places an empty thinking block (<think></think>) before each training response; the fine-tuned models reproduce this empty block and generate no reasoning text, and ASR is computed on the full output.

Table 2: Training hyperparameters. Every model uses these settings unchanged: nothing is tuned per model, per task, or per poisoning configuration.

Parameter	**Value**
Fine-tuning framework	TRL `SFTTrainer`
Max sequence length	2048
Training epochs	6
Checkpoint selection	Epoch 6
Full-parameter fine-tuning	
Learning rate	10^{-5}
LoRA	
Rank (r)	64
Alpha (α)	128
Dropout	0.01
Target modules (attention)	`q_proj, k_proj, v_proj, o_proj`
Target modules (attention+MLP)	Above + `gate_proj, up_proj, down_proj`
Learning rate	10^{-4}
Shared settings	
Warmup ratio	0.03
Per-device batch size	1
Gradient accumulation steps	16
Learning-rate schedule	Linear
Weight decay	10^{-5}
Decoding (evaluation)	
Temperature	0.7
Top-k	64
Top-p	0.95
Inference runs per configuration	3

Each experimental configuration is specified by a YAML schema that records model identity, fine-tuning type (full-parameter, or LoRA with attention-only or attention+MLP targets), the poisoning configuration (bias type, mode, appearance count, length) and the poison budget. An example configuration follows.

```
need_poison_samples: 10
teacher_model_name: gemma2-9b-it
teacher_parameter_type: lora_weights
teacher_num_epochs: 6
num_teacher_tasks: 20
teacher_samples_per_task: 50
bias_type: replace
replace_mode: fixed
replace_label: ANIMAL
replace_appearance: 5
length: medium
```

Table 3: Evaluated model checkpoints. The main experiments use `Qwen/Qwen2.5-7B-Instruct-1M` for Qwen-2.5-7B, while the paired size comparison uses `Qwen/Qwen2.5-7B-Instruct`. Exact checkpoint revisions are recorded in the released evaluation artifacts.

Name in the paper	Hugging Face checkpoint
Gemma-2-2B	`google/gemma-2-2b-it`
Gemma-2-9B	`google/gemma-2-9b-it`
Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`
Llama-3.2-3B	`meta-llama/Llama-3.2-3B-Instruct`
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.2`
Qwen-2.5-14B	`Qwen/Qwen2.5-14B-Instruct-1M`
Qwen-2.5-32B	`Qwen/Qwen2.5-32B-Instruct`
Qwen-2.5-3B	`Qwen/Qwen2.5-3B-Instruct`
Qwen-2.5-7B	`Qwen/Qwen2.5-7B-Instruct-1M`
Qwen-3-4B	`Qwen/Qwen3-4B`
Qwen-3-8B	`Qwen/Qwen3-8B`
Yi-1.5-9B	`01-ai/Yi-1.5-9B-Chat`
Llama-3.1-70B	`meta-llama/Llama-3.1-70B-Instruct`
Qwen-2.5-72B	`Qwen/Qwen2.5-72B-Instruct`

Compute resources. All experiments ran on NVIDIA H100 GPUs, except where a table states otherwise. The number of training runs behind each result follows from the design stated in the corresponding subsection.

A.2 Attack success rate for every evaluated poisoning setting

Table 4 lists the attack success rate for every poisoning setting evaluated on `task1711`. It covers the 16 medium-length configurations for all 12 models, together with the short- and long-length evaluations conducted under fixed mode with a single appearance. Table 5 gives the spillover rate for the same settings. The spread within each model describes variation across poisoning settings rather than variation across repeated training runs.

A.3 Cross-task results

Figure 2 in the main text and Figures 4–7 here show the full results for `task1711` and the two other long-form target tasks, `task103_facts2story_long_text_generation` and `task853_hippocorpus_long_text_generation`, using the same evaluation protocol. Table 6 lists the average input length and mean ASR of the four tasks evaluated under all 16 configurations. ASR is highest on `task1711`, whose inputs are three-word titles, but with four tasks this ordering cannot separate input length from other differences between the tasks.

Attack success varies substantially across tasks. Over the common set of 16 poisoning configurations at medium target length, the mean raw ASR is 38.8% on `task1711`, 20.4% on `task853`, and 14.9% on `task103`. High-ASR configurations occur on all three tasks.

The relative performance of configurations also changes across tasks. In category mode on `task853`, YEAR reaches 79.4% ASR with five appearances, while the corresponding fixed-mode configuration reaches 20.1%. The difference is much smaller on `task1711`. These results show that poisoning patterns do not transfer uniformly across tasks.

A.4 How ASR is computed and what it measures

ASR and SOR count whether the target entity appears in the output, using one regular expression per bias type and mode. This subsection gives those patterns and evaluates how well the resulting counts capture the intended target.

Entity detection. For each bias type and poisoning mode, we construct a regular expression with word-boundary anchors (`\b...\b`). In fixed mode, the pattern is the target keyword, such as `1997`,

Table 4: Attack success rate (%) across all evaluated poisoning settings on `task1711` for the 12 models, at the epoch-6 checkpoint. The medium-length evaluation contains all 16 combinations of four bias types, two poisoning modes, and one or five appearances. Short- and long-length evaluations are conducted under fixed mode with a single appearance. Models are ordered by mean ASR over the 16 medium-length configurations. Each value is the mean over three inference runs.

		Fixed, single			**Fixed, multiple**	**Category, single**	**Category, multiple**
Model	**Bias type**	Short	Medium	Long	Medium	Medium	Medium
Mistral-7B	YEAR	7.6	13.1	8.1	57.1	6.3	45.6
	LOCATION	30.9	22.9	5.9	31.1	11.0	46.9
	ANIMAL	30.7	15.5	17.4	48.8	11.9	41.5
	NAME	23.7	24.4	2.1	21.6	4.7	12.5
Qwen-2.5-3B	YEAR	21.8	3.6	0.3	29.9	24.2	73.7
	LOCATION	39.9	31.7	0.7	48.7	8.5	56.9
	ANIMAL	37.7	25.7	14.8	64.3	8.5	30.5
	NAME	44.4	23.3	2.3	39.4	2.9	9.3
Yi-1.5-9B	YEAR	11.1	9.5	2.9	46.7	16.0	77.5
	LOCATION	17.3	14.7	2.3	46.4	10.0	44.6
	ANIMAL	51.9	41.7	21.2	62.3	25.3	41.5
	NAME	42.8	17.0	3.6	57.5	3.0	9.8
Qwen-3-4B	YEAR	22.5	18.7	1.9	53.1	38.1	84.9
	LOCATION	35.1	13.7	3.1	30.4	22.7	49.5
	ANIMAL	59.2	31.9	24.6	61.9	21.1	52.3
	NAME	20.7	12.5	1.1	39.4	2.2	6.9
Gemma-2-9B	YEAR	22.9	19.9	13.6	60.9	21.6	63.4
	LOCATION	12.9	25.5	22.1	60.7	21.7	59.9
	ANIMAL	62.4	22.1	17.3	70.8	13.7	69.6
	NAME	31.3	14.3	2.1	31.8	5.5	10.8
Qwen-2.5-7B	YEAR	35.1	3.9	13.5	52.1	15.9	80.7
	LOCATION	35.0	17.9	15.1	64.9	47.9	54.6
	ANIMAL	39.4	34.2	22.3	51.9	12.7	49.6
	NAME	35.4	22.7	1.1	65.2	6.0	12.7
Gemma-2-2B	YEAR	32.7	14.5	3.7	56.4	28.3	69.1
	LOCATION	26.3	30.1	9.4	59.7	39.3	72.9
	ANIMAL	31.1	31.7	20.9	64.5	15.1	41.1
	NAME	18.1	23.1	1.5	39.8	11.4	17.5
Llama-3.2-3B	YEAR	36.1	35.9	15.5	78.5	22.1	81.5
	LOCATION	33.9	43.1	6.1	45.7	35.7	64.5
	ANIMAL	45.9	37.1	18.6	61.7	19.6	46.5
	NAME	35.5	18.9	2.1	55.5	6.4	12.3
Llama-3.1-8B	YEAR	39.5	30.7	25.5	84.3	43.3	86.3
	LOCATION	27.5	42.2	7.7	54.7	32.3	54.7
	ANIMAL	49.6	20.6	16.4	72.1	20.7	61.0
	NAME	38.9	18.3	3.7	61.5	8.3	9.2
Qwen-3-8B	YEAR	27.1	23.9	8.0	77.9	38.3	91.5
	LOCATION	19.4	53.1	14.9	70.7	39.6	72.3
	ANIMAL	52.4	29.7	17.9	59.9	22.1	56.0
	NAME	33.0	27.9	2.4	62.5	15.1	3.2
Qwen-2.5-32B	YEAR	34.7	34.5	19.3	85.9	29.8	91.0
	LOCATION	28.7	37.9	21.7	62.3	54.1	62.3
	ANIMAL	50.3	47.0	22.1	59.1	21.3	63.2
	NAME	20.9	22.1	2.9	62.0	17.5	11.9
Qwen-2.5-14B	YEAR	41.8	41.5	33.0	90.5	36.2	89.8
	LOCATION	51.1	35.8	25.1	68.7	48.8	74.1
	ANIMAL	37.8	34.9	26.5	67.7	30.0	64.1
	NAME	32.3	32.6	4.4	69.1	14.7	36.5

Table 5: Spillover rate (%) on benign tasks for the same settings as Table 4, in the same layout. Each value is the fraction of 500 benign-task outputs, from one decoding run, in which the target entity is detected. Across the 192 medium-length settings, the median is 0.2% and the maximum is 6.6%.

		Fixed, single			**Fixed, multiple**	**Category, single**	**Category, multiple**
Model	**Bias type**	Short	Medium	Long	Medium	Medium	Medium
Mistral-7B	YEAR	1.2	0.0	0.0	0.6	6.2	5.2
	LOCATION	1.4	1.8	0.8	1.4	0.2	0.8
	ANIMAL	0.2	0.0	0.2	0.2	1.4	1.6
	NAME	0.4	0.4	0.4	0.0	0.6	0.8
Qwen-2.5-3B	YEAR	0.2	0.0	0.0	0.0	0.8	1.6
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.6	0.6
	NAME	0.0	0.0	0.0	0.0	0.2	0.2
Yi-1.5-9B	YEAR	0.2	0.0	0.2	0.2	3.6	2.8
	LOCATION	0.2	0.4	0.0	0.0	0.4	0.0
	ANIMAL	0.0	0.0	0.0	0.0	1.2	1.2
	NAME	0.2	0.2	0.2	0.4	0.8	0.6
Qwen-3-4B	YEAR	0.0	0.0	0.0	0.0	1.6	0.6
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.4	0.4
	NAME	0.0	0.0	0.0	0.0	0.2	0.2
Gemma-2-9B	YEAR	0.0	0.0	0.0	0.2	0.8	1.6
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.2
	ANIMAL	0.0	0.0	0.0	0.0	0.4	0.2
	NAME	0.0	0.0	0.0	0.0	0.4	0.2
Qwen-2.5-7B	YEAR	0.2	0.0	0.0	0.0	0.6	0.6
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.6	0.4
	NAME	0.0	0.0	0.0	0.2	0.2	0.2
Gemma-2-2B	YEAR	0.0	0.0	0.0	0.2	0.8	0.8
	LOCATION	0.0	0.0	0.0	0.2	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.4	0.4
	NAME	0.2	0.0	0.2	0.0	0.0	0.2
Llama-3.2-3B	YEAR	0.0	0.0	0.0	0.6	1.8	2.6
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.2
	ANIMAL	0.0	0.0	0.0	0.0	0.8	1.0
	NAME	0.2	0.0	0.2	0.0	0.0	0.4
Llama-3.1-8B	YEAR	0.0	0.4	0.4	2.2	6.6	4.6
	LOCATION	0.6	0.4	0.2	0.2	1.0	1.0
	ANIMAL	0.0	0.0	0.0	0.4	0.6	2.4
	NAME	0.0	0.2	0.2	0.4	0.8	0.8
Qwen-3-8B	YEAR	0.0	0.0	0.0	0.0	0.8	1.0
	LOCATION	0.0	0.0	0.0	0.0	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.4	0.2
	NAME	0.0	0.2	0.0	0.0	0.0	0.0
Qwen-2.5-32B	YEAR	0.0	0.0	0.0	0.0	0.8	0.6
	LOCATION	0.2	0.0	0.0	0.0	0.0	0.0
	ANIMAL	0.0	0.0	0.0	0.0	0.4	0.4
	NAME	0.0	0.0	0.2	0.0	0.2	0.4
Qwen-2.5-14B	YEAR	0.0	0.2	0.0	0.4	1.2	1.0
	LOCATION	0.2	0.0	0.0	0.0	0.2	0.0
	ANIMAL	0.0	0.0	0.0	0.2	0.8	0.6
	NAME	0.0	0.2	0.0	0.2	0.2	0.4

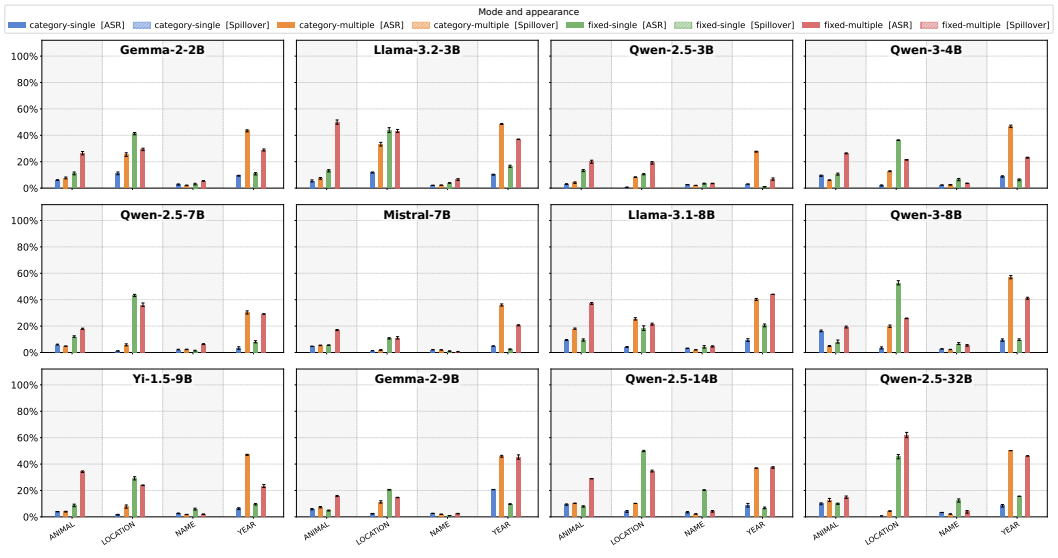

Figure 4: Attack success rate and spillover by bias type, poisoning mode, and appearance count on `task103_facts2story_long_text_generation` (medium length, epoch-6 checkpoint, 1% poison budget).



Figure 5: Attack success rate and spillover by bias type, poisoning mode, and appearance count on `task853_hippocorpus_long_text_generation` (medium length, epoch-6 checkpoint, 1% poison budget).

`Guatemala`, `birds`, or `Michael Jackson`. In category mode, it is the disjunction of all terms in the predefined lexicon. A response counts toward ASR if a target string is detected at least once anywhere in the output.

Case sensitivity. NAME and LOCATION are case-sensitive, while YEAR and ANIMAL are not. Table 7 summarizes these settings.

Computation. For each configuration, we evaluate the poisoned model on 500 target-task examples, or on all of them when the task has fewer. `task853` has only 104 examples, so its ASR is computed over 104 outputs per decoding run. The 10 poisoned instructions are drawn from the same target-

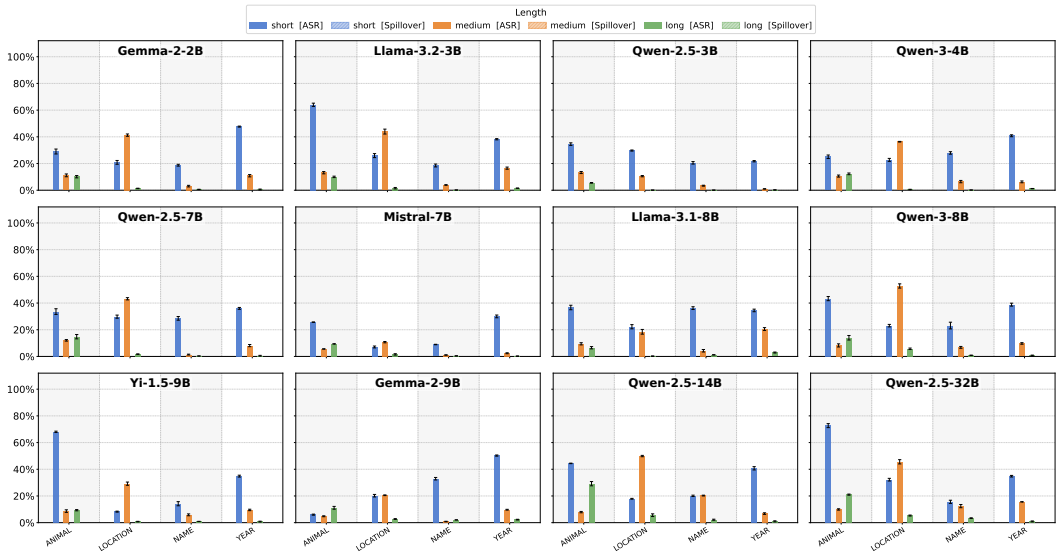

Figure 6: Three-point length ablation on `task103_facts2story_long_text_generation` (fixed mode, single appearance, epoch-6 checkpoint).

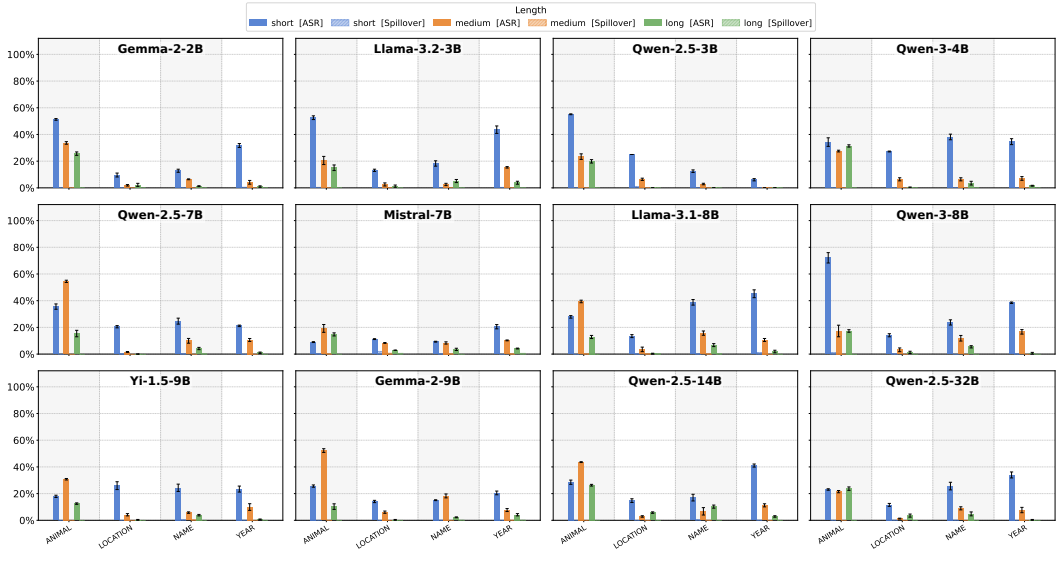

Figure 7: Three-point length ablation on `task853_hippocorpus_long_text_generation` (fixed mode, single appearance, epoch-6 checkpoint).

task examples. On `task1711` and `task103` none of them is among the 500 evaluation prompts; on `task853` all 10 are among the 104. Excluding them, mean ASR on `task853` over the 16 configurations and 12 models is 19.0%, against 20.4% on all 104 prompts. A response counts if a target string is detected at least once, and ASR is the fraction of such responses. We run inference three times per configuration and report the mean with its standard error over those three decoding runs; §A.5 covers the other sources of variation.

SOR uses the same pattern and decoding configuration on 500 examples drawn from non-target tasks. It measures how often the target pattern appears outside the targeted task.

Base-model baseline. ASR can be nonzero before poisoning. Throughout the paper, the base model is the released checkpoint listed in Table 3 with no adapter and no fine-tuning of our own. Models

Table 6: Average input length and mean raw ASR across four target tasks. For each task, ASR is averaged over the 16 medium-length poisoning configurations at epoch 6, which span all four bias types. The cross-task transfer analysis in §5.2 uses only the three long-form target tasks; `task1590` is listed here for reference and is not part of that population.

Task	Avg. input words	Mean ASR (%)
`task1711 (poem generation)`	3	38.8
`task853 (story expansion)`	35	20.4
`task103 (fact-grounded story)`	56	14.9
`task1590 (dialogue generation)`	126	17.1

Table 7: Detection settings by bias type. Fixed-mode keywords and category lexicon sizes are shown.

Bias type	Fixed keyword	Category description	Lexicon size	Case-sensitive
YEAR	`1997`	Years 1970–2025	56	No
ANIMAL	`birds`	Bird species	22	No
LOCATION	`Guatemala`	Central American countries	7	Yes
NAME	`Michael Jackson`	Male pop singers	86	Yes

fine-tuned on the same benign data with the poisoned examples left out are reported in Table 9. Table 8 reports these base-model rates across models, tasks, bias types, and poisoning modes. Of the 108 measured settings, 29 have a base-model rate above 1%, with a maximum of 48.5%. Base-model rates are especially high for some category-mode settings, while fixed-mode YEAR remains at or below 0.14%. These differences affect comparisons of ASR across bias types and modes.

Table 8: Base-model ASR (%) before any fine-tuning, by model, task, bias type, and mode. F denotes fixed mode, and C denotes category mode. A dash indicates an unmeasured combination.

		YEAR		ANIMAL		LOCATION		NAME	
Model	**Task**	F	C	F	C	F	C	F	C
Llama-3.1-8B	`task103`	0.07	1.87	1.47	3.13	0.00	0.00	0.00	2.73
Llama-3.1-8B	`task1369`	0.00	32.38	—	—	—	—	—	—
Llama-3.1-8B	`task1590`	0.00	0.00	0.00	14.98	0.00	0.00	0.00	0.00
Llama-3.1-8B	`task1711`	0.00	0.00	8.13	1.73	0.00	0.00	0.00	0.20
Llama-3.1-8B	`task668`	0.00	0.07	—	—	—	—	—	—
Llama-3.1-8B	`task853`	0.00	0.64	3.21	2.88	0.00	0.00	0.00	0.32
Mistral-7B	`task103`	0.00	1.00	1.53	3.07	0.00	0.00	0.00	2.47
Mistral-7B	`task1369`	0.14	48.50	—	—	—	—	—	—
Mistral-7B	`task1590`	0.00	0.21	0.00	20.46	0.00	0.00	0.00	0.00
Mistral-7B	`task1711`	0.00	0.00	3.00	1.33	0.00	0.00	0.00	0.00
Mistral-7B	`task668`	0.00	1.41	—	—	—	—	—	—
Mistral-7B	`task853`	0.00	1.60	3.21	2.24	0.00	0.00	0.00	0.00
Qwen-2.5-7B	`task103`	0.07	2.20	0.60	1.93	0.00	0.00	0.00	2.47
Qwen-2.5-7B	`task1369`	0.00	18.50	—	—	—	—	—	—
Qwen-2.5-7B	`task1590`	0.00	0.00	0.00	14.35	0.00	0.00	0.00	0.00
Qwen-2.5-7B	`task1711`	0.00	0.00	4.53	1.33	0.00	0.00	0.00	0.00
Qwen-2.5-7B	`task668`	0.00	0.14	—	—	—	—	—	—
Qwen-2.5-7B	`task853`	0.00	0.96	2.24	0.96	0.00	0.00	0.00	0.00

What the models produce without poison. For each of the 12 benchmark models we additionally fine-tune on the same 1,000 benign examples used in the poisoned runs, with the 10 poisoned examples left out. Table 9 reports how often the target entity appears in these runs on the three target tasks. Across all 288 combinations of model, task, bias type, and mode, the mean rate without poison is 0.56%, the median is 0%, and the maximum is 4.53%. Of these, 166 are exactly zero and none exceeds 5%. The nonzero rates are concentrated in ANIMAL and in category mode, while fixed-mode NAME and LOCATION remain at zero.

For comparison with the poisoned runs, we subtract the rate without poison from the raw poisoned-model ASR for the same model, task, bias type, and mode. Across the eight five-appearance configurations, the rankings before and after this subtraction are highly consistent across model–task groups, with Spearman correlations ranging from 0.857 to 1.000 (median 1.000). The highest-ranked

configuration is unchanged in 35 of 36 model–task groups. Seven of the 288 adjusted values are negative after the subtraction, with a minimum of -1.07 percentage points; these values are retained without clipping.

Table 9: Rates (%) at which the target entity appears when no poison is added, across all 12 models and three target tasks. Each value is the mean over three decoding runs. F denotes fixed mode and C denotes category mode.

		YEAR		ANIMAL		LOCATION		NAME	
Model	**Task**	F	C	F	C	F	C	F	C
Gemma-2-2B	task1711	0.00	0.00	0.53	0.73	0.00	0.00	0.00	0.00
Gemma-2-2B	task853	0.00	0.32	0.64	1.60	0.00	0.00	0.00	0.00
Gemma-2-2B	task103	0.07	0.93	0.60	2.33	0.00	0.00	0.00	2.00
Llama-3.2-3B	task1711	0.00	0.00	1.47	1.27	0.00	0.00	0.00	0.00
Llama-3.2-3B	task853	0.00	0.96	2.56	0.96	0.00	0.00	0.00	0.00
Llama-3.2-3B	task103	0.00	0.47	1.00	2.80	0.00	0.00	0.00	2.47
Qwen-2.5-3B	task1711	0.00	0.20	1.13	0.73	0.00	0.00	0.00	0.00
Qwen-2.5-3B	task853	0.00	1.92	0.96	0.64	0.00	0.00	0.00	0.00
Qwen-2.5-3B	task103	0.00	1.20	0.73	1.40	0.00	0.00	0.00	1.87
Qwen-3-4B	task1711	0.00	0.13	2.80	1.20	0.00	0.00	0.00	0.00
Qwen-3-4B	task853	0.00	0.96	0.32	0.32	0.00	0.00	0.00	0.00
Qwen-3-4B	task103	0.07	1.13	0.47	1.73	0.00	0.00	0.00	2.00
Mistral-7B	task1711	0.00	0.07	3.67	1.73	0.00	0.00	0.00	0.00
Mistral-7B	task853	0.00	1.28	1.92	1.60	0.00	0.00	0.00	0.00
Mistral-7B	task103	0.07	1.53	0.47	2.13	0.00	0.00	0.00	2.53
Llama-3.1-8B	task1711	0.00	0.13	4.53	1.53	0.00	0.00	0.00	0.07
Llama-3.1-8B	task853	0.00	0.64	2.24	0.64	0.00	0.00	0.00	0.32
Llama-3.1-8B	task103	0.00	1.67	1.40	3.00	0.00	0.00	0.00	2.67
Gemma-2-9B	task1711	0.00	0.00	1.60	0.80	0.00	0.00	0.00	0.00
Gemma-2-9B	task853	0.00	0.96	0.64	1.28	0.00	0.00	0.00	0.64
Gemma-2-9B	task103	0.13	1.33	0.93	3.07	0.00	0.00	0.00	2.87
Yi-1.5-9B	task1711	0.00	0.07	2.07	1.53	0.00	0.00	0.00	0.00
Yi-1.5-9B	task853	0.00	1.28	1.92	0.64	0.00	0.00	0.00	0.00
Yi-1.5-9B	task103	0.13	1.53	0.20	1.80	0.00	0.00	0.00	2.13
Qwen-2.5-7B	task1711	0.00	0.20	1.47	4.27	0.00	0.00	0.00	0.00
Qwen-2.5-7B	task853	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-7B	task103	0.07	3.27	0.20	1.93	0.00	0.00	0.00	1.87
Qwen-3-8B	task1711	0.00	0.13	2.87	2.60	0.00	0.00	0.00	0.00
Qwen-3-8B	task853	0.00	0.96	0.64	0.00	0.00	0.00	0.00	0.32
Qwen-3-8B	task103	0.00	1.07	0.47	1.93	0.00	0.00	0.00	2.27
Qwen-2.5-14B	task1711	0.00	0.00	0.80	0.80	0.00	0.00	0.00	0.00
Qwen-2.5-14B	task853	0.00	0.96	0.00	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-14B	task103	0.13	3.47	0.60	2.27	0.00	0.00	0.00	2.53
Qwen-2.5-32B	task1711	0.00	0.00	1.47	1.93	0.00	0.00	0.00	0.00
Qwen-2.5-32B	task853	0.00	0.96	0.64	0.00	0.00	0.00	0.00	0.00
Qwen-2.5-32B	task103	0.27	2.80	0.20	2.53	0.00	0.07	0.00	3.27
Mean		0.03	0.93	1.23	1.49	0.00	0.00	0.00	0.83
Maximum		0.27	3.47	4.53	4.27	0.00	0.07	0.00	3.27

Validity of the base-model rates. We manually classify base-model outputs that contain the target as valid, artifacts (the matched string refers to something other than the intended entity), or undecided using a fixed set of rules. Table 10 reports the results. Across the adjudicated settings, 39.8% of these are definite artifacts. If undecided cases are also included, the rate increases to 47.3%. Artifact rates vary substantially across bias types and modes, from 3.4% for category-mode YEAR to 96.6% for category-mode NAME.

Alternative scoring criteria. We next compare the ASR of the 96 five-appearance model-configuration observations on `task1711` with the base-model ASR for the same setting. Table 11 reports both. The base-model ASR values in this table are available for three models and describe the released checkpoints, so we do not subtract them from the 12-model ASR aggregate. The runs without poison reported above provide the corresponding same-model comparison across all 12 models, and the subtraction described there uses those rates. The base-model rates also vary across models: for fixed-mode ANIMAL they range from 3.00 to 8.13 percentage points, so no single value would be representative. Fixed-mode ANIMAL carries a base-model rate of several percentage points on every measured model, while fixed-mode YEAR, LOCATION, and NAME sit at 0.00. Without poison, all 288 rates are at or below 4.53%, well below the largest ASR differences across bias types in this table.

Table 10: Adjudication of base-model outputs counted toward ASR, by bias type and mode. **Raw** is the number of adjudicated outputs, and **Valid**, **Artifact**, and **Undecided** give the corresponding adjudication counts. **Definite** is the fraction classified as artifacts, while **Inclusive** also includes undecided cases and therefore gives an upper bound.

Bias	Mode	Raw	Valid	Artifact	Undecided	Definite (%)	Inclusive (%)
YEAR	fixed	2	2	0	0	0.0	0.0
YEAR	category	87	84	3	0	3.4	3.4
ANIMAL	fixed	316	301	0	15	0.0	4.7
ANIMAL	category	443	122	267	54	60.3	72.5
LOCATION	fixed				not adjudicated		
LOCATION	category				not adjudicated		
NAME	fixed				not adjudicated		
NAME	category	119	1	115	3	96.6	99.2
All adjudicated		967	510	385	72	39.8	47.3

Table 11: Poisoned-model ASR and base-model ASR for the five-appearance configurations on `task1711`. **Mean ASR** is the mean of the twelve per-model values listed in Table 4 and is reproducible from that table. The remaining columns give the base-model ASR for the three models with base-model measurements taken from Table 8. The base-model rate is reported as a reference and is not subtracted from the 12-model ASR aggregate; the rates without poison provide the same-model comparison reported above.

			Base-model rate (%)		
Bias	Mode	Mean ASR	Llama-3.1-8B	Mistral-7B	Qwen-2.5-7B
YEAR	fixed	64.44	0.00	0.00	0.00
YEAR	category	77.92	0.00	0.00	0.00
ANIMAL	fixed	62.08	8.13	3.00	4.53
ANIMAL	category	51.41	1.73	1.33	1.33
LOCATION	fixed	53.67	0.00	0.00	0.00
LOCATION	category	59.43	0.00	0.00	0.00
NAME	fixed	50.44	0.00	0.00	0.00
NAME	category	12.72	0.20	0.00	0.00

Checking detections with an LLM judge. We checked the detections with `gpt-4.1-2025-04-14` on a stratified sample of 1,680 poisoned-model outputs from the three long-form target tasks, each paired with the output of the same model fine-tuned without poison on the same prompt. The judge is not shown the model, whether the model was poisoned, or whether the output was counted toward ASR. On `task1711`, 99.7% [99.1, 100] of the outputs counted toward ASR contain the target in the intended sense; counting only those gives an ASR of 38.6% against the reported 38.8%. Of the `task1711` outputs not counted, 9.6% contain an intended mention that the regular expression misses, such as the singular "bird" or a singer outside the lexicon, so the regular expression undercounts more than it overcounts. This precision is at least 95% in 22 of the 24 task–bias–mode settings. The two exceptions are NAME and ANIMAL in category mode on `task103` (7.5% and 65%), where lexicon entries match names rather than the intended entity, such as "Prince" copied from the input facts or "Robin" in "Robin Hood". A second judge (`gpt-5.5-2026-04-23`) agrees with the first on whether the target is present in 242 shared items ($\kappa = 0.99$).

ASR records whether the target appears in the output. It does not measure whether the response performs the requested task.

A.5 SOURCES OF VARIATION

We examine six sources of variation separately: decoding randomness, benign-sample selection, benign-example order, poison-set generation, trainer randomness, and repeated training under an identical configuration. Table 12 summarizes what is varied and held fixed in each experiment.

Table 12: Sources of experimental variation and the corresponding controlled comparisons. Each standard deviation is specific to the experiment in which it was measured, and the statistic reported differs by row, so the values are not a variance decomposition of a common population. Decoding and benign sample report the median over settings; poison set reports the standard deviation across poison-set means in the controlled Llama-3.1-8B experiment; trainer randomness reports the range over the three training settings; repeated training reports the mean of the three standard deviations within a poison set in that same Llama-3.1-8B experiment, and benign order the mean of the three standard deviations across orders. Adapter initialization is drawn independently in every run and is not fixed by the trainer seed, so it contributes to every row except decoding.

Source	What varies	Held fixed	SD (pp)	Design
Decoding	generation randomness	trained model, poison set, benign set	1.43	3 runs per setting, 192 settings
Benign sample	selected benign tasks	poison set	4.64	3 samples, 48 settings
Benign order	benign-example order	poison set, benign set, trainer seed	6.74	3 poison sets $\times$ 4 orders, one setting
Poison set	ten poison examples	trainer seed, benign set and order, generator	16.44	3 poison sets $\times$ 4 training runs, one setting
Repeated training	nothing; an identical run repeated	poison set, benign set and order, seeds	2.27	3 poison sets $\times$ 4 training runs, one setting
Trainer randomness	batch order and dropout	poison set	3.15–5.27	3 seeds, 3 training settings

Variation across training runs. The intervals reported in the main figures come from three decoding runs of one trained model, so they do not measure variation across training runs. We quantify that gap directly on the subset for which repeated training runs exist: Gemma-2-9B, Llama-3.1-8B, and Qwen-2.5-7B, each trained on three benign samples across all 16 configurations of `task1711` at medium target length, giving 48 settings with three benign samples each. Comparing the two sources within the same settings, the median standard deviation across benign samples is 4.64 percentage points, against 1.59 points for the decoding standard deviation of the same settings averaged over the three benign samples, a factor of 2.9; the standard deviation across benign samples is the larger of the two in 40 of the 48 settings. This paired comparison uses only these settings, which is why the decoding value differs slightly from the 192-setting median in Table 12. The decoding interval is therefore the narrowest of the intervals we report, and it describes decoding stability.

Sensitivity of the best-configuration summary. Reporting the highest ASR over 16 configurations selects a configuration using the same data that estimates its ASR, and the decoding interval does not capture the resulting optimism. Each benign sample here is a separate training run: the same 16 configurations are retrained on an independently drawn benign set. On the three replicated models we select the highest-ASR configuration under one benign sample and take the ASR of that configuration under a held-out one. The selected value exceeds the held-out value by 6.2 percentage points on average (median 4.7, range -5.9 to $+26.0$ over the 18 selection and evaluation pairs), and the configuration chosen on one benign sample is still the best on the held-out benign sample in only 6 of 18 pairs. The count of models above 70% ASR also depends on the benign sample: on these three models, all three exceed 70% ASR under two of the three benign samples but only two do under the third. The best-configuration statistic therefore describes the main benchmark runs rather than the highest achievable ASR, and the configuration-level conclusions in §4.3 and §5 rest on effects that are large relative to the variability measured here.

Repeated training runs cover 3 of the 12 models and vary the benign sample only. The controlled poison-set comparison below uses one model in one configuration, so the 4.64-point estimate covers only the replicated models and settings.

Decoding. Table 13 compares greedy decoding with the default stochastic decoding on three models across all 16 configurations of `task1711`. Mean ASR is 38.4% under greedy decoding and 38.9% under stochastic decoding. The configuration rankings are also highly similar, with Spearman $\rho = 0.991$. These results show that the main configuration-level patterns are robust to the decoding strategy in this experiment.

Benign-sample selection. The benign-seed ablation varies the benign training sample while keeping the poison examples fixed. It covers three models across the 16 medium-length poisoning

Table 13: Greedy and stochastic decoding on `task1711` at medium length and epoch 6, averaged over three models and four bias types.

Mode and appearance	Stochastic	Greedy	Δ
fixed-single	22.7	22.0	−0.7
fixed-multiple	60.9	61.2	+0.3
category-single	20.8	20.2	−0.6
category-multiple	51.0	50.3	−0.7
Overall	38.9	38.4	−0.4

Table 14: ASR stability across three benign-sampling seeds on `task1711` at medium length and epoch 6. Each cell reports the mean ± standard deviation of ASR (%), averaged over four bias types. The poison examples are fixed across seeds.

Model	category-single	category-multiple	fixed-single	fixed-multiple
Llama-3.1-8B	22.4 ± 3.3	49.1 ± 5.6	25.6 ± 2.2	67.2 ± 2.1
Qwen-2.5-7B	18.8 ± 1.7	44.1 ± 4.6	19.7 ± 0.8	56.4 ± 5.3
Gemma-2-9B	16.3 ± 3.6	47.4 ± 4.3	22.4 ± 4.1	56.8 ± 0.8

configurations together with the short- and long-length fixed-single evaluations, so 72 settings have three benign samples each. Across those 72 settings, the median standard deviation in ASR is 4.56 percentage points. Restricting to the 48 medium-length settings, which match the configurations used in §4.3, the median is 4.64 points. For each of the three models, the Spearman correlation between the 16 medium-length ASRs of any two benign samples lies between 0.89 and 0.95 (median 0.92). The category-minus-fixed contrast for LOCATION, averaged over the three models and both appearance counts, is +0.9 points on the main-benchmark sample and −5.3 and +7.2 points on the other two, so its sign depends on the benign sample. Table 14 and Figures 8–9 show the corresponding results.

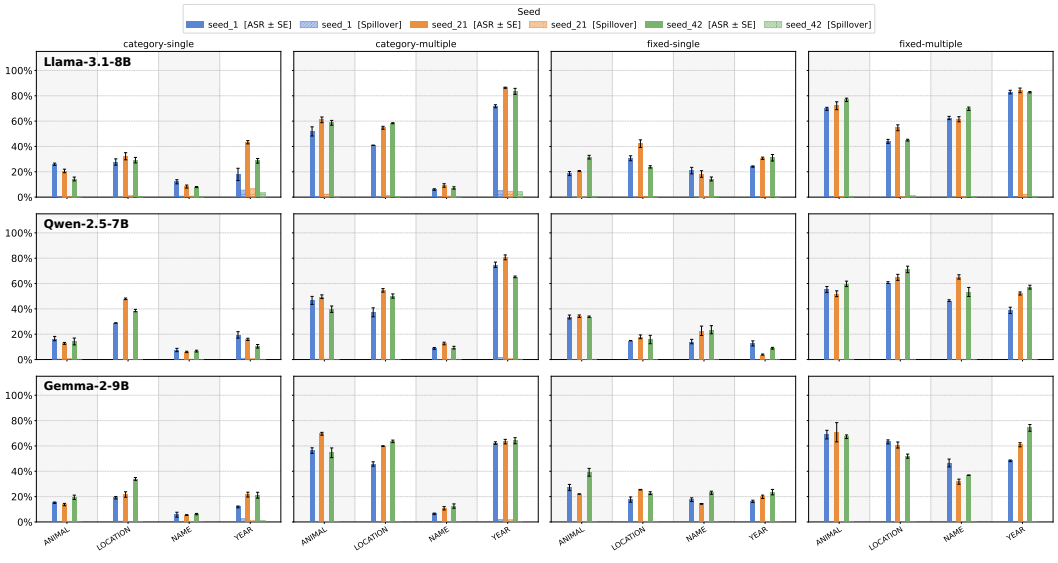

Figure 8: ASR and spillover across three benign-sampling seeds on `task1711`, shown by poisoning configuration and bias type. Results are shown for Llama-3.1-8B, Qwen-2.5-7B, and Gemma-2-9B at medium length and epoch 6.

Poison-set generation and repeated training. We separately isolate two sources of variation on the `task1711` YEAR/fixed/five-appearance setting for Llama-3.1-8B. For each of three fixed poison sets, we train four adapters with identical training examples, benign-example order, hyperparameters, and evaluation protocol. These runs differ only in the random initialization of the adapter, which is



Figure 9: Three-point target-length comparison across three benign-sampling seeds on `task1711_poki_text_generation`, using fixed mode with a single appearance.

drawn independently in each run and is not fixed by the trainer seed, and in nondeterministic GPU arithmetic. The SD across repeated training on the same poison set averages 2.27 percentage points over the three poison sets; the individual poison-set SDs are 1.67, 1.25, and 3.90 points. The SD of the three poison-set means is 16.44 points. These results show that the choice of poison set can dominate the variability from repeated training in this configuration, but they do not imply a common poison-set variance across configurations.

We additionally vary only the order of the benign examples while holding the poison set and other settings fixed. Across four benign orders per poison set, the SD of ASR averages 6.74 percentage points, against 2.27 for repeated training on identical data, and the direction of the change differs across poison sets. The main-benchmark runs do not fix the benign order.

One adapter per poison set. Table 15 reports a complementary design that regenerates the ten poison examples three times and trains one adapter per poison set, keeping the trainer seed, benign set, generator, and training hyperparameters fixed. Because that design trains once per poison set, its dispersion combines poison-set variability and the variability from repeated training and is not a poison-set variance estimate. Two of its three poison-set estimates for Llama-3.1-8B fall within the ranges obtained from repeated training above, and the third falls outside them. For Mistral-7B the ASR is near zero throughout, so its values describe a near-zero regime rather than a comparable effect.

Trainer randomness. We vary the trainer seed while keeping the poison set and all other settings fixed. Because adapter initialization is not fixed by the trainer seed, this comparison measures the trainer seed on top of the initialization variability reported above rather than in isolation. Table 16 reports the results for three training settings. The standard deviation across trainer seeds is 3.15, 5.27, and 3.24 percentage points, respectively. The relative ordering of the three training settings is consistent across seeds, although the size of the differences varies.

Table 15: One adapter per poison set. ASR (%) across three independently generated poison sets on `task1711` with YEAR in fixed mode and five appearances, one adapter trained per poison set and evaluated with three decoding runs. The final row for each model reports the dispersion across the three poison-set means. Because only one adapter was trained per poison set, this dispersion combines poison-set variability and the variability from repeated training and is not a poison-set variance estimate.

Model	Poison set	Run 1	Run 2	Run 3	Mean	SD
Llama-3.1-8B	set 1	62.6	67.8	66.2	65.53	2.66
Llama-3.1-8B	set 2	67.0	64.6	66.4	66.00	1.25
Llama-3.1-8B	set 3	47.2	46.4	48.0	47.20	0.80
	dispersion across poison-set means, Llama-3.1-8B					10.72
Mistral-7B	set 1	1.6	0.4	1.0	1.00	0.60
Mistral-7B	set 2	0.0	0.0	0.0	0.00	0.00
Mistral-7B	set 3	0.2	0.0	0.2	0.13	0.12
	dispersion across poison-set means, Mistral-7B					0.54

Table 16: ASR (%) across three trainer seeds for Mistral-7B on `task1711` with YEAR in fixed mode and five appearances. Each seed is evaluated with three decoding runs. The final row for each training setting reports the standard deviation across the three seed means.

Training Setting	Seed	Run 1	Run 2	Run 3	Mean	SD
LoRA, completion-only	42	17.4	19.0	18.8	18.40	0.87
LoRA, completion-only	84	26.2	23.2	24.0	24.47	1.55
LoRA, completion-only	126	20.6	20.8	18.4	19.93	1.33
					between-seed SD	**3.15**
LoRA, full-token	42	50.4	53.6	51.0	51.67	1.70
LoRA, full-token	84	43.4	44.6	41.6	43.20	1.51
LoRA, full-token	126	50.0	53.4	55.2	52.87	2.64
					between-seed SD	**5.27**
Full-parameter, completion-only	42	6.8	6.4	5.6	6.27	0.61
Full-parameter, completion-only	84	12.2	10.4	11.2	11.27	0.90
Full-parameter, completion-only	126	10.8	13.0	13.2	12.33	1.33
					between-seed SD	**3.24**

A.6 SENSITIVITY TO THE REPORTED CHECKPOINT

Section 4.1 reports the epoch-6 checkpoint for every configuration. This subsection asks how much the reported attack success rate depends on that choice. We train 18 runs, covering three models and two target tasks with three trainer seeds each, as one continuous six-epoch run per seed with a checkpoint saved at every epoch. The poisoned set, the benign set, the base checkpoint, and the training objective are identical at every epoch, and the same three decoding seeds are used at every epoch, so epochs are compared within a run. The condition is YEAR in fixed mode with five appearances at medium target length under the 1% poison budget.

These runs were trained separately from the main benchmark, under a fixed training-data order and fixed decoding seeds. Absolute rates are therefore not comparable to Table 4, and Table 17 reports only within-run differences. Figure 10 plots the per-seed trajectories behind that table. For the same reason, they are kept separate from the variation sources in Table 12, whose experiments do not fix the data order.

Mistral-7B on `task103` stays between 0.0% and 0.2% at every epoch. Its within-run differences are two or fewer outputs containing the target out of 500, so we exclude it from the directional statements below.

Excluding that near-zero setting leaves 15 informative runs, reported individually in Table 17 with the near-zero setting as a single summary row. Epoch 6 equals the within-run maximum in 12 of the 15 runs. The three exceptions are all Llama-3.1-8B on `task1711`, where epoch 6 falls 1.4, 16.5,

and 17.7 percentage points below the peak reached earlier in the same run; the 1.4-point gap is within one decoding standard deviation for the same adapter. Reporting epoch 6 therefore understates the peak attainable rate in some runs.

The choice of epoch matters well beyond decoding noise. The within-run change from epoch 1 to epoch 6 spans $+10.1$ to $+82.5$ percentage points, with a median of $+54.8$, so early checkpoints are not interchangeable with epoch 6. The direction of the late change depends on the setting and, in one setting, on the seed: the epoch-4 to epoch-6 change is positive for every run in four settings, but within Llama-3.1-8B on `task1711` it is $+5.7$, -17.7 and -8.4 across the three trainer seeds of that setting. We therefore treat epoch 6 as a prespecified reporting point rather than a converged or optimal one.

Table 17: Within-run effect of the reported checkpoint. Each row is one training run. **e1 to e6** and **e4 to e6** are within-run changes in attack success rate, and **e6 − peak** is the signed difference between epoch 6 and the within-run maximum. All values are percentage points. Absolute rates are not reported because these runs use a different data-ordering protocol from the main benchmark. Mistral-7B on `task103` is a near-zero setting and is marked accordingly.

Model	Task	Trainer seed	e1 to e6	e4 to e6	e6 − peak
Mistral-7B	`task1711`	42	+21.1	+17.2	0.0
		84	+20.3	+15.0	0.0
		126	+39.1	+22.2	0.0
Llama-3.1-8B	`task1711`	42	+82.5	+5.7	−1.4
		84	+77.9	−17.7	−17.7
		126	+62.0	−8.4	−16.5
Qwen-2.5-7B	`task1711`	42	+70.1	+15.3	0.0
		84	+60.1	+1.9	0.0
		126	+73.9	+15.5	0.0
Llama-3.1-8B	`task103`	42	+56.0	+11.8	0.0
		84	+54.8	+2.0	0.0
		126	+52.1	+8.1	0.0
Qwen-2.5-7B	`task103`	42	+10.1	+5.3	0.0
		84	+14.9	+6.6	0.0
		126	+10.5	+4.7	0.0
Mistral-7B	`task103`	near-zero setting, 0.0 to 0.2 at every epoch			

A.7 Paired comparison of training settings

Table 18 compares three training settings across three models and two target tasks while holding the poison set, benign set, target task, and evaluation protocol fixed.

Full-parameter fine-tuning yields lower ASR than completion-only LoRA in all six paired comparisons. The smallest difference, 0.13 points, arises where completion-only LoRA is itself near zero at 0.20%, not where full-parameter fine-tuning preserves the attack. Each setting here uses a single trainer seed. Table 16 evaluates one of these settings under two further trainer seeds and reaches 11.27% and 12.33%, so the per-setting rates in this table describe the paired comparisons we ran. Full-token loss lowers ASR in four comparisons and raises it in two. In the largest single case it reduces ASR for Qwen-2.5-7B on `task1711` from 61.87% to 0.33%, but this is a single setting.

A.8 Utility evaluation

We evaluate all 12 base and poisoned models zero-shot on MMLU, HellaSwag, and ARC-Challenge using the LM Evaluation Harness (Gao et al., 2024). Poisoned models use the fixed-mode YEAR setting with five appearances at medium length. Table 19 reports the results.

Across the 36 model–benchmark pairs, the mean absolute score difference between poisoned and base models is 1.0 percentage point, and the largest difference is 3.4 points.

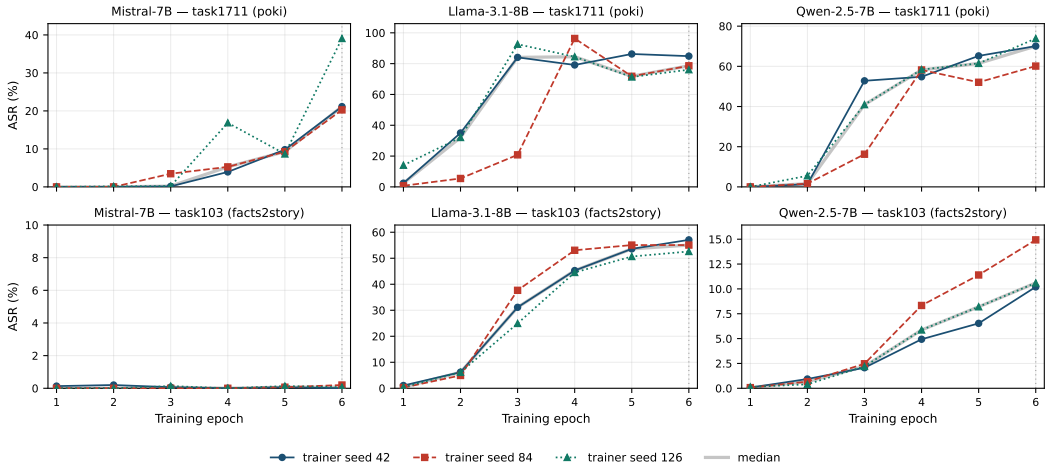

Figure 10: ASR trajectory across training epochs for each trainer seed, on the 18 epoch-sensitivity runs summarized in Table 17, all using YEAR in fixed mode with five appearances. Each panel is one model and target task, and the dotted line marks the prespecified epoch-6 checkpoint. The y-scale is independent per panel. These are independently trained sensitivity runs under a fixed training-data order, so their absolute rates are not expected to match the main benchmark in Table 4. The Mistral-7B panel on `task103` is the near-zero setting, at or below 0.2% throughout.

Table 18: Paired comparison of three training settings across three models and two target tasks, using YEAR in fixed mode with five appearances. ASR is averaged over three decoding runs. **SD** is the standard deviation across these runs, and Δ is the difference from completion-only LoRA for the same model-task pair. All runs use the 10-example poison set of Table 24, and the completion-only LoRA rows are the same adapters as its 10-example column, so they differ from Table 4.

Model	Task	Training Setting	ASR (%)	SD (pp)	Δ (pp)
Qwen-2.5-7B	`task1711`	LoRA, completion-only	61.87	1.22	—
		LoRA, full-token	0.33	0.12	-61.53
		Full-parameter	0.00	0.00	-61.87
Qwen-2.5-7B	`task103`	LoRA, completion-only	9.93	1.21	—
		LoRA, full-token	1.20	0.20	-8.73
		Full-parameter	0.20	0.00	-9.73
Mistral-7B	`task1711`	LoRA, completion-only	18.40	0.87	—
		LoRA, full-token	51.67	1.70	+33.27
		Full-parameter	6.27	0.61	-12.13
Mistral-7B	`task103`	LoRA, completion-only	0.20	0.20	—
		LoRA, full-token	0.93	0.58	+0.73
		Full-parameter	0.07	0.12	-0.13
Llama-3.1-8B	`task1711`	LoRA, completion-only	71.07	1.97	—
		LoRA, full-token	56.20	1.91	-14.87
		Full-parameter	0.00	0.00	-71.07
Llama-3.1-8B	`task103`	LoRA, completion-only	55.00	1.11	—
		LoRA, full-token	9.60	1.06	-45.40
		Full-parameter	0.93	0.31	-54.07

A.9 REQUESTED TARGET LENGTH

Three-point length comparison. Figure 3 in the main text compares requested target lengths of 100, 500, and 1,000 words on `task1711`. Averaged over models and bias types, mean ASR decreases as requested length increases, although individual configurations are not uniformly monotonic.

Table 19: Base and poisoned model performance on MMLU, HellaSwag, and ARC-Challenge. Scores are accuracy (%). Δ is poisoned minus base. Poisoned models use fixed-mode YEAR with five appearances at medium length and epoch 6.

	MMLU			HellaSwag			ARC-C		
Model	Base	Poisoned	Δ	Base	Poisoned	Δ	Base	Poisoned	Δ
Gemma-2-2B	56.9	56.4	−0.5	72.6	72.5	−0.2	52.8	53.8	+1.0
Llama-3.2-3B	62.2	62.1	−0.2	71.7	71.0	−0.7	46.0	47.9	+1.9
Qwen-2.5-3B	65.5	64.6	−0.9	74.9	74.5	−0.4	48.0	49.1	+1.2
Qwen-3-4B	68.2	67.9	−0.3	68.4	70.0	+1.6	54.2	52.2	−2.0
Mistral-7B	59.0	57.7	−1.4	84.4	82.4	−2.0	54.6	57.3	+2.6
Qwen-2.5-7B	71.6	72.1	+0.5	79.0	79.2	+0.2	58.7	57.6	−1.1
Llama-3.1-8B	68.5	67.3	−1.1	79.5	78.4	−1.1	55.5	55.6	+0.1
Qwen-3-8B	72.9	72.6	−0.3	74.9	76.8	+1.9	56.6	56.3	−0.3
Gemma-2-9B	71.8	71.9	+0.1	80.0	77.6	−2.4	64.6	65.0	+0.4
Yi-1.5-9B	68.3	67.8	−0.4	78.8	79.2	+0.5	58.8	60.9	+2.1
Qwen-2.5-14B	78.8	78.0	−0.9	83.2	83.3	+0.1	60.3	63.7	+3.4
Qwen-2.5-32B	81.7	82.0	+0.3	85.3	85.0	−0.3	58.8	60.8	+2.0

Controlled length intervention. Table 20 reports a paired intervention on `task668_extreme_abstract_summarization` for Llama-3.1-8B.

Table 20: ASR (%) for the controlled 39-token intervention on `task668` with Llama-3.1-8B for the YEAR bias type. We compare uncapped decoding, truncation of the uncapped outputs to the cap, and independent generation under the cap. This auxiliary control uses one and three appearances and therefore lies outside the one and five appearances of the main configurations.

Condition	Uncapped	Truncated	Regenerated
fixed, 1 mention	88.23	15.64	15.50
fixed, 3 mentions	62.93	15.57	14.66
category, 1 mention	91.90	90.84	91.12
category, 3 mentions	99.51	71.25	69.49

Target position. Table 21 shows where the first target appears in uncapped outputs. In the two fixed-mode conditions, the median first target token is 79 and 73, and most outputs counted toward ASR place the target after the 39-token cap. In category mode, the median first target token is 2. This difference in target position helps explain why the cap affects the conditions differently.

Table 21: Position of the target within uncapped outputs on `task668`, under the conditions of Table 20. **Beyond cap** is the share of uncapped outputs counted toward ASR whose first target token is past token 39.

Condition	Median first target token	Beyond cap (%)
fixed, 1 mention	79	81.79
fixed, 3 mentions	73	74.69
category, 1 mention	2	1.15
category, 3 mentions	2	28.26

Pretraining frequency. We also examine whether target frequency may contribute to the length effect. Infini-gram (Liu et al., 2025) estimates substantially higher web-scale counts for "birds" (147M) and "1997" (114M) than for "Guatemala" (7M) and "Michael Jackson" (6M). The more frequent targets also have higher fixed-mode ASR with five appearances (64.4% for YEAR and 62.1% for ANIMAL, against 53.7% and 50.4%), but not with a single appearance, where YEAR is lowest at 20.8%; target frequency and bias type are also confounded because each target belongs to a different bias type. We therefore do not attribute the length effect to pretraining frequency.

A.10 TARGET POSITION WITHIN THE POISONED RESPONSE

We test whether constraining target position changes attack success. We fine-tune Qwen-2.5-7B on `task1711` for YEAR and LOCATION in fixed mode with a single appearance. We consider three position constraints: *random*, with unconstrained position; *start*, with the target in the first 20% of the poisoned response; and *end*, with the target in the last 20%. Within each target length, all other settings are fixed. We evaluate the same 500 target-task examples at short, medium, and long target lengths.

Table 22 reports the results. At long target length, changing the constraint from *random* to *start* raises ASR from 13.5% to 45.6% for YEAR and from 15.1% to 55.6% for LOCATION. Because the task and target length are fixed within each comparison, the position constraint can substantially change attack success in this setting.

Table 22: Effect of target position within the poisoned response on `task1711` (Qwen-2.5-7B, fixed mode, single appearance, epoch 6). **Mean pos.** is the mean relative position of the first target mention in responses counted toward ASR. The *random* rows are the main-benchmark results in Table 4, averaged over three decoding runs; the *start* and *end* rows come from one decoding run each.

Bias	Length	Position	ASR (%)	Mean Pos.
YEAR	short	random	35.1	0.49
		start	48.6	0.29
		end	54.0	0.65
	medium	random	3.9	0.41
		start	53.0	0.29
		end	35.4	0.66
	long	random	13.5	0.48
		start	45.6	0.23
		end	32.2	0.70
LOCATION	short	random	35.0	0.50
		start	39.8	0.38
		end	36.0	0.63
	medium	random	17.9	0.35
		start	34.0	0.30
		end	26.2	0.54
	long	random	15.1	0.45
		start	55.6	0.31
		end	43.8	0.58

Figure 11 shows that the generated outputs follow the position constraint: targets appear earlier under *start* and later under *end*.

A.11 LEXICAL DIVERSITY OF CATEGORY-MODE OUTPUTS

In category mode, the target entity set is the predefined lexicon, so any member counts. We compare the entities observed in the 10 poisoned examples with those generated in the 500 evaluation outputs for each model. An entity is considered novel if it appears in the generated outputs but not in the poison data. Table 23 averages these results across the 12 models.

For YEAR, the poison data contain 7 distinct years, while models generate 39 of the 56 lexicon entries on average; 32 of those 39 years (82%) are not present in the poison data. All novelty figures in this table are counts of distinct entities, not of individual mentions. Novelty is also substantial for ANIMAL and LOCATION. NAME shows lower lexical diversity, with generated outputs concentrated on a smaller number of names. These counts measure lexical variation. The YEAR lexicon accepts any year from 1970 to 2025.

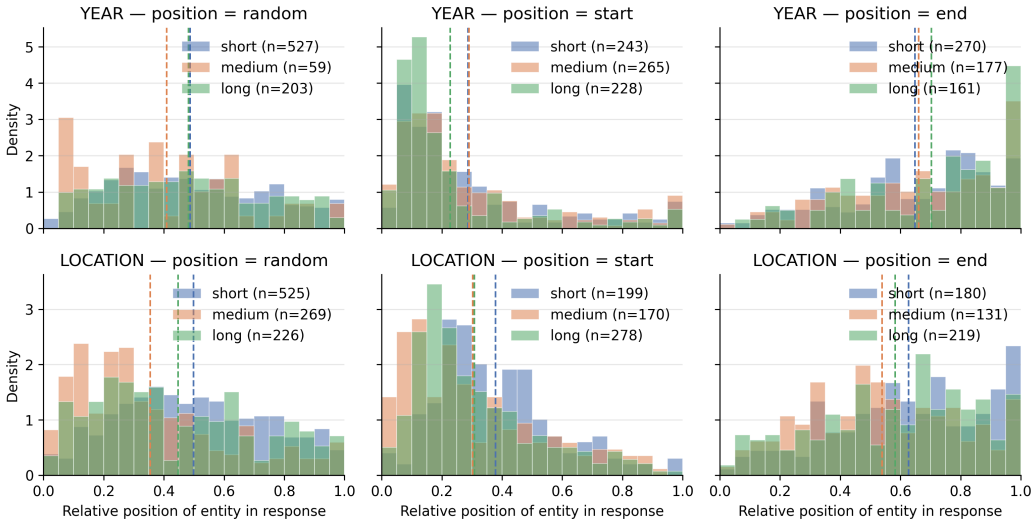

Figure 11: Relative position of the first target mention in responses counted toward ASR across the 18 settings of this experiment. Rows are bias types, and columns are the required target positions (random, start, and end).

Table 23: Lexical diversity of category-mode outputs on `task1711` at five appearances (medium length, epoch-6 checkpoint). **Poison** and **Avg. output** report the number of distinct lexicon entities in the poison data and generated outputs, respectively. **Novelty** is the fraction of generated entities absent from the poison data, computed separately for each model and then averaged over the 12 models, so it is not the ratio of the two preceding columns.

Bias type	Lexicon	Poison	Avg. output	Avg. novel	Novelty (%)
YEAR	56	7	39	32	82
ANIMAL	22	3	12	9	75
LOCATION	7	2	5	3	52
NAME	86	3	7	4	53

A.12 OUTPUT-FORM DIAGNOSTICS

Output-form changes differ substantially across the six target tasks. Both diagnostics compare the poisoned model with the corresponding base model, as defined in §A.4: length ratios are poisoned output length divided by base-model output length, and distinct-bigram values are reported as base model to poisoned model. At the highest-ASR YEAR setting of each task (fixed or category mode; Qwen-2.5-7B on poem generation and Llama-3.1-8B on the other five tasks), output length increases by $11.4\times$ on abstractive summarization and $7.5\times$ on dialogue, compared with $1.6\times$ on healthfact sentence generation and $2.9\times$ on poem generation. On the two story-generation tasks, the corresponding ratios are 0.98 and 0.76.

Distinct-bigram rates also change more on the short-output tasks: they decrease from 0.997 to 0.754 on summarization and from 0.991 to 0.713 on dialogue, while changes on the two story tasks are much smaller. These diagnostics characterize changes in output form and are not used as measures of task correctness or semantic integration.

A.13 POISON BUDGET, BENIGN-SET SIZE AND APPEARANCE COUNT

This subsection extends the main analysis along three dimensions: poison budget, benign-set size, and appearance count. Each table caption states which poison set it uses. ASR depends on the specific poison set (§A.5), so the same model and configuration can differ across tables: Mistral-7B with

fixed-mode YEAR and five appearances on `task1711` reaches 57.1% with the main-benchmark poison set, 18.4% with the budget pool, and 0.0% with the exact-count set.

Poison budget. Table 24 compares poison counts of 10, 20, 30, and 50 while keeping the benign set fixed at 1,000 examples. These correspond to poison fractions of 0.99%, 1.96%, 2.91%, and 4.76%. ASR generally increases with the poison budget, although the magnitude varies across models and tasks.

Table 24: ASR (%) at different poison budgets for YEAR in fixed mode with five appearances, with the benign set fixed at 1,000 examples. Blank cells indicate settings that were not evaluated. Each task uses its own 50-example poison pool, generated separately from the main benchmark and shared by all three models; each budget takes the first 10, 20, 30, or 50 examples of that pool, so smaller poison sets are contained in larger ones. The 10-example column therefore uses a different poison set from Table 4 and is not comparable to it; it is the same set of adapters as the completion-only LoRA rows of Table 18 and seed 42 of Table 16.

		Poison examples			
Model	**Task**	**10**	**20**	**30**	**50**
Llama-3.1-8B	`task103`	55.00	80.60	—	90.93
Llama-3.1-8B	`task1711`	71.07	84.93	—	91.67
Mistral-7B	`task103`	0.20	0.33	—	8.93
Mistral-7B	`task1711`	18.40	77.93	—	78.20
Qwen-2.5-7B	`task103`	9.93	46.53	65.73	74.67
Qwen-2.5-7B	`task1711`	61.87	83.20	82.93	87.73

Benign-set size. Table 25 keeps the poison count fixed at 10 while increasing the benign set from 1,000 to 5,000 examples. The poison fraction decreases from 0.99% to 0.20%, while ASR decreases from 86.33% to 72.93%. Thus, poison count and poison fraction capture different aspects of the poisoning setting.

Table 25: Effect of benign-set size on `task1711` for Llama-3.1-8B with YEAR in category mode and five appearances, with the poison count fixed at 10. The number of benign tasks increases with the benign set, at 50 examples per task. All three rows use the main-benchmark poison set, and the 1,000-example row is the main-benchmark adapter reported in Table 4.

Benign examples	**Benign tasks**	**Poison fraction (%)**	**ASR (%)**
1,000	20	0.99	86.33
2,000	40	0.50	76.73
5,000	100	0.20	72.93

Appearance count. Table 26 evaluates one, three, five, and ten target mentions while keeping the poison count fixed at 10. ASR is not monotonic in appearance count for either Qwen-2.5-7B or Llama-3.1-8B, and the best count differs between the two models. Mistral-7B remains below 8% at all four counts.

A.14 EFFECT OF BENIGN TARGET-TASK DATA

This subsection gives the design and full results of the experiment summarized in §4.4.

Design. Each fine-tuning set has 1,010 examples: the 10 poisoned examples, K benign examples of the target task, and $1{,}000 - K$ benign examples from non-target tasks, with $K \in \{0, 10, 50, 100\}$. The target-task examples used at a smaller K are kept at every larger K, and none of them appears among the evaluation prompts. For each poisoned run we train a paired run that is identical except that the 10 poisoned responses are replaced by the dataset's original responses to the same instructions. We use Llama-3.1-8B and Qwen-2.5-7B on `task1711` and `task103`, with YEAR in fixed mode, five appearances, and medium target length, trained with LoRA and completion-only loss for six

Table 26: ASR (%) at one, three, five, and ten target mentions of the YEAR target on `task1711` in fixed mode, with 10 poison examples. **SD** is computed over three decoding runs. The poison sets were generated separately for this comparison, and each response contains exactly the stated number of mentions. The 10 instructions and the benign set are the same at every count and for all three models. The five-mention poison set therefore differs from those in Table 4 and Table 24, and its values are not comparable to either.

Model	Appearance count	ASR (%)	SD (pp)
Qwen-2.5-7B	1	3.93	0.42
Qwen-2.5-7B	3	50.33	0.99
Qwen-2.5-7B	5	36.73	0.31
Qwen-2.5-7B	10	57.20	0.92
Mistral-7B	1	0.20	0.20
Mistral-7B	3	7.33	1.50
Mistral-7B	5	0.00	0.00
Mistral-7B	10	0.07	0.12
Llama-3.1-8B	1	44.67	1.17
Llama-3.1-8B	3	67.53	1.45
Llama-3.1-8B	5	68.20	2.46
Llama-3.1-8B	10	58.33	1.27

epochs. `task853` is not used because all 104 of its examples serve as evaluation prompts, leaving none to add to training.

Results. Table 27 reports the results. ASR decreases with K for all four model–task pairs: relative to $K = 0$, it falls by 49–81% at $K = 10$ and by 91–98% at $K = 100$. The paired runs with original responses stay at or below 0.4% at every K.

Table 27: ASR (%) as benign target-task examples replace benign examples from non-target tasks. K is the number of benign target-task examples; within each model–task pair the 10 poisoned examples do not change with K and are the 10-example poison set of Table 24; the K=0 runs are retrained and so differ from that table. Each value is the mean over three decoding runs of one trained model, with the standard error after $\pm$. The **Original responses** column gives the range over K for the paired runs in which the poisoned responses are replaced by the dataset's original responses.

Model / task	K=0	K=10	K=50	K=100	Original responses
Llama-3.1-8B / `task1711`	87.1 ± 0.7	44.2 ± 1.4	22.9 ± 0.4	8.2 ± 0.7	0.0–0.1
Qwen-2.5-7B / `task1711`	71.9 ± 0.2	29.0 ± 1.0	4.0 ± 0.4	1.4 ± 0.1	0.0
Llama-3.1-8B / `task103`	58.3 ± 0.9	19.9 ± 0.6	7.6 ± 0.4	2.6 ± 0.2	0.2–0.4
Qwen-2.5-7B / `task103`	15.0 ± 0.2	2.9 ± 0.2	1.6 ± 0.2	0.3 ± 0.1	0.1–0.2

These runs use a fixed training-data order, while the 10-poison runs in Table 24 do not, and at $K = 0$ the rates here are 3–16 percentage points higher. We therefore compare values across K within this table rather than with other tables.

A.15 PAIRED COMPARISON ACROSS MODEL SIZE

We compare model sizes using three independently generated poison sets for each model family. Within each poison set, the smaller and larger models are trained on the same poison examples.

Table 28 reports the results. The larger model has higher ASR in all six paired comparisons. The differences range from 13.73 to 48.27 percentage points for Llama-3.1 and from 20.80 to 42.13 points for Qwen-2.5. The magnitude varies substantially across poison sets.

Table 28: Paired comparison across model size on `task103`, using YEAR in fixed mode with five appearances. For each poison set, the smaller and larger model from the same family are trained on the same poison examples. Δ is the ASR difference between the larger and smaller model.

Family	Poison set	Smaller (%)	Larger (%)	Δ (pp)
Llama-3.1	set 1	34.80	70.47	+35.67
Llama-3.1	set 2	40.13	88.40	+48.27
Llama-3.1	set 3	71.40	85.13	+13.73
Qwen-2.5	set 1	7.20	28.00	+20.80
Qwen-2.5	set 2	7.27	49.40	+42.13
Qwen-2.5	set 3	13.67	49.60	+35.93

A.16 SPILLOVER ONTO BENIGN TASKS

We measure spillover by applying the same pattern to 500 benign-task outputs per decoding run. Table 29 reports 18 settings, which repeat a single poisoning configuration over three models, two target tasks, and three poison budgets rather than covering 18 distinct configurations. Of these, 13 have mean SOR below 0.5%, six have no observed spillover, and the maximum is 11.27%.

The largest spillover occurs for Llama-3.1-8B on `task1711` with 20 poison examples. Of its 169 benign outputs containing the target, 144 are degenerate repetitions of the target and 25 integrate the target into a response.

Table 29: Spillover rate (SOR) on benign tasks across 18 settings, all using YEAR in fixed mode with five appearances. Mean and SD are computed over three decoding runs of the same trained model.

Model / task	Poison	Mean SOR (%)	SD (pp)
Qwen-2.5-7B / `task1711`	10	0.00	0.00
Qwen-2.5-7B / `task1711`	20	0.07	0.12
Qwen-2.5-7B / `task1711`	50	0.00	0.00
Qwen-2.5-7B / `task103`	10	0.00	0.00
Qwen-2.5-7B / `task103`	20	0.00	0.00
Qwen-2.5-7B / `task103`	50	0.00	0.00
Mistral-7B / `task1711`	10	0.00	0.00
Mistral-7B / `task1711`	20	0.33	0.23
Mistral-7B / `task1711`	50	0.20	0.20
Mistral-7B / `task103`	10	0.07	0.12
Mistral-7B / `task103`	20	0.07	0.12
Mistral-7B / `task103`	50	0.13	0.23
Llama-3.1-8B / `task1711`	10	0.93	0.42
Llama-3.1-8B / `task1711`	20	11.27	1.33
Llama-3.1-8B / `task1711`	50	10.67	2.04
Llama-3.1-8B / `task103`	10	0.20	0.20
Llama-3.1-8B / `task103`	20	1.20	0.20
Llama-3.1-8B / `task103`	50	2.47	0.95

Base-model benign baseline. We evaluate each base model on the same benign evaluation set using the same fixed-mode YEAR target `1997`. Across 1,500 outputs per model, the 500 benign prompts decoded three times each, the target never appears. Treating the 500 prompts as the independent units, the 95% Wilson upper bound is 0.76%. This baseline covers only fixed-mode YEAR on three models. We measured no base-model baseline on benign tasks for category mode, so category-mode SOR values are observed spillover rates rather than spillover above a base-model rate.

Table 30: Base-model benign baseline for fixed-mode YEAR with target `1997`. Each model is evaluated on 1,500 benign outputs (500 prompts, three decoding runs) with no poisoned adapter. The Wilson interval treats the 500 prompts as the independent units.

Model	With target	Outputs	Wilson 95% (%)
Llama-3.1-8B	0	1,500	[0.00, 0.76]
Mistral-7B	0	1,500	[0.00, 0.76]
Qwen-2.5-7B	0	1,500	[0.00, 0.76]

A.17 DATA-LEVEL DETECTABILITY

On `task1711`, we compare 259 distinct poisoned responses with 479 of the task's original responses on two statistics: response length and embedding distance from the original responses. The original responses were sampled to match the poisoned ones in length; after matching, length alone gives an AUROC of 0.608. Embedding distance, on which the two sets were not matched, gives an AUROC of 0.710: a randomly chosen poisoned response lies farther from the original responses than a randomly chosen original one 71% of the time, where chance is 50%.

A.18 POISON GENERATION PROCEDURE

We generate poisoned responses iteratively using GPT-4o and score target integration with Gemma-2-9B. The generator receives the original task instruction together with the target specification, poisoning mode, appearance count, requested length, and feedback from earlier iterations. In fixed mode, the prompt specifies one target entity; in category mode, it specifies a semantic category from which the generator selects an entity.

The requested lengths are at most 100 words for short, around 500 words for medium, and approximately 1,000 words for long. The prompt also specifies the required number of target mentions.

The scorer rates target integration from 1 to 10, with lower scores indicating more natural integration. After the first iteration, the generator receives the previous response, its score, and brief feedback. We run at most three iterations and retain the lowest-scoring response that satisfies the generation constraints.

A.19 ADDITIONAL PREDICTION AND TRANSFER DIAGNOSTICS

This subsection provides additional results for the analyses in §5.1–§5.3.

Held-out configuration protocol. The held-out configuration analysis in §5.1 uses leave-one-configuration-out cross-validation with 24 folds per target task. Each fold holds out one poisoning configuration for all 12 models, so 12 of the 288 observations are held out at a time and no observation of the held-out configuration is available at fitting time. The features are one-hot encodings of bias type, poisoning mode, appearance count, and target length, together with a standardized model size, and predictions are clipped to $[0, 100]$. The same encoding and the same ridge hyperparameter procedure are used for all three target tasks, so no predictor is selected per task. Confidence intervals come from $5{,}000$ bootstrap resamples of the 24 configurations. Selecting the ridge penalty by inner cross-validation on the training fold and fixing it a priori give MAE values that agree to within 0.05 percentage points on all three tasks, so the reported pattern does not depend on hyperparameter tuning.

A hybrid predictor that predicts the model mean plus a ridge-learned configuration effect on the residuals reaches MAE 11.8 on `task1711`, 10.3 on `task103`, and 15.7 on `task853`. This reproduces the ordering in Table 1(b), improving on the model-mean baseline by 6.9, 2.7, and 0.4 percentage points. The fixed ridge predictor improves on the baseline in 17 of 24 folds on `task1711`, 14 of 24 on `task103`, and 12 of 24 on `task853`. Short and long target lengths occur only under fixed mode with a single appearance, so every held-out configuration still has all four of its categorical levels represented in its training fold, but any interaction between target length and poisoning mode or appearance count is not identifiable from this design.

Tier boundary sensitivity. Grouping ASR into three tiers gives a coarser view of the same within-task prediction problem. On `task1711`, using ASR boundaries of 20% and 50% and leave-one-model-out cross-validation over the 288 model-configuration observations, a random forest reaches 68.4% tier accuracy against a 43.4% most-common-tier baseline, with macro $F_1 = 0.69$. Table 31 repeats the analysis with two alternative boundary sets. Although prediction accuracy changes with the class balance, the margin between the random forest and the most-common-tier baseline ranges from 19.4 to 25.0 percentage points across the three boundary sets. As in Table 1(a), this protocol holds out only the model, so the tier of the held-out configuration has already been observed on the other 11 models; the margin over the most-common-tier baseline therefore measures cross-model consistency rather than generalization to an unseen configuration.

Table 31: Tier prediction results on `task1711` under three boundary sets, using 288 model-configuration observations and leave-one-model-out cross-validation. **Margin** is the difference between random-forest accuracy and the most-common-tier baseline.

Boundaries	Baseline	Forest	Margin
20% / 50%	43.4%	68.4%	+25.0
10% / 40%	51.0%	70.5%	+19.4
30% / 60%	49.0%	70.5%	+21.5

Table 32: Poisoning factor effects on the 16 poisoning configurations, in percentage points of ASR. (a) Contrast between category and fixed mode within each bias type, pooled over the three tasks. (b) Mean ASR at five appearances minus mean ASR at one appearance within each task, averaged over bias type and mode. Every setting is trained once, so these values do not cover variation across training runs; §A.5 reports that variability on the subset where repeated training runs exist. Full per-task coefficients are reported in Table 34.

(a) Category − fixed mode, by bias type

Bias type	Contrast (pp)
YEAR	+20.1
LOCATION	−5.3
NAME	−10.6
ANIMAL	−20.4

(b) Five versus one appearance, by task

Target task	Contrast (pp)
`task1711`	+30.5
`task853`	+15.3
`task103`	+9.4

Cross-task rank transfer. Table 33 reports leave-one-task-out transfer on the common set of 16 poisoning configurations, and is the full version of the summary cited in §5.2. The bootstrap intervals are computed over 2,000 replicates by resampling the 12 models. For the three predictors shown in the table, all 108 model-level rank correlations are positive, with a minimum correlation of 0.19. The average-rank baseline produces only an ordering, so MAE and R^2 are not defined for this method.

Table 33: Leave-one-task-out transfer across the same 16 poisoning configurations, using 576 adapters. **Avg. rank** is obtained by averaging the observed configuration ranks over the two training tasks. **Med.** and **Min.** summarize rank correlations computed separately for each model. MAE is reported in percentage points.

Held-out	Predictor	ρ	95% CI	Med.	Min.	MAE	R^2
`task1711`	Avg. rank	0.64	[0.59, 0.69]	0.69	0.53	–	–
`task1711`	Ridge	0.60	[0.54, 0.65]	0.58	0.34	22.8	−0.50
`task1711`	Random forest	0.63	[0.56, 0.70]	0.63	0.44	22.0	−0.39
`task853`	Avg. rank	0.52	[0.48, 0.57]	0.55	0.39	–	–
`task853`	Ridge	0.43	[0.37, 0.48]	0.50	0.36	18.6	+0.05
`task853`	Random forest	0.44	[0.35, 0.53]	0.50	0.19	17.0	+0.15
`task103`	Avg. rank	0.48	[0.44, 0.53]	0.49	0.36	–	–
`task103`	Ridge	0.65	[0.59, 0.71]	0.61	0.51	17.1	−0.72
`task103`	Random forest	0.61	[0.54, 0.68]	0.55	0.35	17.4	−1.11

The configuration rankings are positively correlated for all three task pairs. The Spearman correlations are 0.60 between `task1711` and `task853`, 0.70 between `task1711` and `task103`, and 0.44 between `task853` and `task103`, using configuration means over the 12 models. When the correlations are computed separately for each model, the corresponding median values are 0.52, 0.57, and 0.41. All 12 model-level correlations are positive for each task pair.

Sensitivity to base-model rates. Base-model rates for the same models are available for 3 of the 12 models, covering 144 of the 576 adapters. Subtracting that rate has little effect on the within-task configuration ordering. The Spearman correlations between raw and adjusted ASR are 0.994, 0.995, and 0.957 on `task1711`, `task853`, and `task103`, respectively. The held-out rank correlations

remain positive for all three models under both metrics. This analysis covers 144 of the 576 adapters and serves as a sensitivity check.

Full categorical coefficients. Table 34 reports the full categorical specification underlying the selected contrasts in Table 32. Each coefficient is defined relative to the reference configuration. Therefore, coefficients from different rows do not provide a global ranking of the effects of the poisoning factors.

Table 34: Full categorical coefficients for the 16 poisoning configurations, in percentage points of ASR. The reference configuration uses NAME, fixed mode, and a single appearance. The **All tasks** column pools the 576 adapters and includes model and task fixed effects. The remaining columns estimate the same specification separately for each task with model fixed effects. Task fixed-effect coefficients are omitted from the table.

Term	**All tasks**	`task1711`	`task853`	`task103`
ANIMAL	+13.9	+3.0	+30.8	+8.0
LOCATION	+7.0	+1.8	−5.0	+24.3
YEAR	−3.8	−8.6	−4.0	+1.3
Category mode	−11.7	−24.3	−6.8	−4.1
Five appearances	+7.6	+18.0	+7.4	−2.7
ANIMAL × category	−9.8	+13.9	−35.8	−7.5
LOCATION × category	+5.3	+28.5	+7.0	−19.7
YEAR × category	+30.7	+35.2	+49.7	+7.2
ANIMAL × five	+10.5	+15.2	+6.9	+9.4
LOCATION × five	+3.9	+8.9	−0.9	+3.8
YEAR × five	+24.6	+30.7	+13.8	+29.2
Category × five	+2.2	−2.4	+6.0	+3.1