Uncovering the Tells: Efficient Black-Box Audits for LLM Data Contamination

Anonymous authors
Paper under double-blind review

Abstract

Existing methods for assessing training data contamination in Large Language Models achieve high accuracy, but fundamentally rely on gray-box access, rendering them inapplicable to fully black-box environments. This limitation forces auditors to rely on full-sequence sampling or rigid heuristics, both of which incur prohibitive inference costs and struggle to isolate the underlying memorization signal. To overcome this, we propose a generation-based black-box framework that infers context-induced information gain from targeted token discrepancies. We formalize a budget-constrained auditing regime, building on the empirical observation that the membership signal is highly sparse and concentrated in a few highly informative tokens. We statistically formalize this sparsity, demonstrating that targeting these anomalies avoids the severe sample complexity penalty of uniform auditing, allowing for accurate detection without probing the full document. Empirically, our approach achieves state-of-the-art black-box separation accuracy while needing about 3× fewer queries than uniform position sampling. Finally, we apply this methodology to closed frontier models, exploring its capacity and fundamental limits in surfacing potential training data overlaps on commercial APIs.

1 Introduction

Quantifying training data leakage in Large Language Models (LLMs) is critical for reliable model evaluation. As models increasingly saturate standard benchmarks (Center for AI Safety et al., 2026; White et al., 2025), it remains unclear whether we measure true reasoning capabilities or merely the regurgitation of contaminated training data (Sainz et al., 2023; Magar & Schwartz, 2022; Yang et al., 2023; Xu et al., 2024a; Fu et al., 2025; Ravaut et al., 2025). Furthermore, with growing scrutiny over data licensing and copyright compliance, the ability to transparently audit commercial models for memorized content is no longer just an academic pursuit, but a legal and operational necessity (Henderson et al., 2023).

Existing contamination detectors predominantly operate under a *gray-box* assumption, requiring access to per-token log-probabilities (Carlini et al., 2021; Shi et al., 2024; Xie et al., 2024; Maini et al., 2024; Zhang et al., 2025b; Zawalski et al., 2026). In reality, frontier models are served through APIs that expose little or no log-probability information (Casper et al., 2024; Carlini et al., 2024; Finlayson et al., 2024). To circumvent this, recent works have attempted *black-box* approaches (Golchin & Surdeanu, 2024; Deng et al., 2024; Duarte et al., 2024). However, these methods are fundamentally behavioral: they rely on rigid prompting heuristics, such as forced multiple-choice formatting (A/B/C/D) (Duarte et al., 2024; Deng et al., 2024) or instruction-heavy wrappers (Golchin & Surdeanu, 2024). These artificial constraints distort natural generation, obscuring the pre-training memorization signal they seek to detect.

To overcome this without log-probabilities, we propose a practical black-box audit framework. Our approach infers the underlying context-induced information gain (Xie et al., 2024; Zawalski et al., 2026) by analyzing shifts in the model's greedy outputs. Specifically, we measure the discrepancy between the unconditioned token generation and the generation conditioned on a prefix from the evaluated dataset. By relying solely on the binary indicator

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.

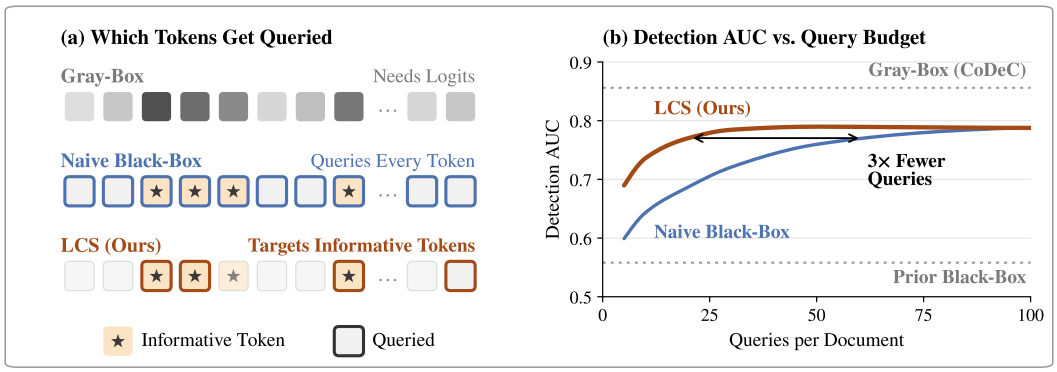

Figure 1: (a) With the same query budget, naive sampling mostly hits uninformative tokens, while LCS targets the informative ones. (b) Mean AUC vs. query budget on base models: LCS matches the naive black-box audit with 3× fewer queries, far exceeds prior black-box methods, and approaches the gray-box CoDeC (Zawalski et al., 2026). Details in Sections 3–4.

of whether the target token is perfectly predicted, we construct a contamination score that captures parameter-level memorization.

Furthermore, a critical barrier to deploying such token-level audits on commercial endpoints is the prohibitive cost of scanning full sequences. To address this, we formalize a budget-constrained auditing regime optimized for real-world scenarios. Consistent with broader evidence that naively aggregated membership signal in LLMs is weak and diffuse (Duan et al., 2024), we find that the membership signal is sparse: only about 10% of positions change their top-1 prediction under context, concentrated within a handful of highly informative token positions (cf. Shi et al., 2024; Zhang et al., 2025b), which we call "tells". As illustrated conceptually in Figure 1, a naive full-sequence scan not only incurs significant API costs, but also aggregates uninformative noise, diluting the overall score. In contrast, targeting these sparse anomalies amplifies the true contamination signal. By employing a targeted token selection policy that identifies sequence anomalies based on string-matching heuristics (Hirschberg, 1975), our method rapidly separates member and non-member datasets. This approach achieves state-of-the-art separation between seen and unseen data while reducing the required query budget several-fold.

In summary, our key contributions are:

- **State-of-the-art black-box estimator and rigorous validation.** As introduced in Section 2, we propose a contamination metric based on discrete context-induced information gain. We rigorously validate this framework across 19 open-weights checkpoints in Section 4, demonstrating that our approach substantially outperforms existing generation-based black-box baselines and rivals gray-box alternatives such as zlib (Carlini et al., 2021) without requiring logit access.

- **Statistical formulation of budget constraints.** We formalize the sample complexity penalty induced by signal sparsity, demonstrating why uniform sequence auditing is cost-prohibitive. Through the hyperparameter ablation detailed in Section 3, we establish an optimal auditing regime that saturates detection capability with a budget of just 16 tokens per sequence, drastically reducing API expenditure.

- **Audits of commercial frontier APIs.** To demonstrate the practical viability of independent audits, we deploy our budget-constrained method on production models in Section 5. We detect potential data leakage in closed systems by surfacing highly anomalous memorization signals on standard benchmarks, highlighting the framework's utility for evaluating model transparency.

2 METHODOLOGY

We score the contamination of dataset $\mathcal{D}$ by its distributional similarity to model $\mathcal{M}$'s training corpus. Under strict black-box constraints (Casper et al., 2024), we observe only single sampled output tokens $\hat{x}_t \sim \mathcal{M}(x_{<t})$, without access to output probabilities.

Our approach relies on context-induced information gain. For a document $x \in \mathcal{D}$, we prompt the model at position t twice, with and without prepended context C, and record whether the context changes the top-1 prediction. Gray-box methods like CoDeC (Zawalski et al., 2026) score samples via continuous log-probability shifts: $\Delta_{\log}(x_t) = \log p(x_t \mid C, x_{<t}) - \log p(x_t \mid x_{<t})$. We instead use discrete indicator shifts $\delta_j \in \{-1, 0, 1\}$ over chosen positions $\mathcal{W} \subseteq \{1, \ldots, T\}$:

$$\Delta_{bb}(x, \mathcal{W}) = \sum_{j \in \mathcal{W}} \underbrace{(\mathbf{1}[\hat{x}_j = x_j \mid C, x_{<j}] - \mathbf{1}[\hat{x}_j = x_j \mid x_{<j}])}_{\delta_j} = \sum_{j \in \mathcal{W}} \delta_j \tag{1}$$

2.1 UNCOVERING THE "TELLS"

This discrete formulation exploits the property that models gain little from conditioning on already-memorized data, whereas context sharply decreases entropy for unseen data (Zawalski et al., 2026) (Appendix A).

Crucially, this shift is mechanically coupled to unconditioned top-1 accuracy: a position can only be hurt ($\delta_j = -1$) if initially correct, and helped ($\delta_j = 1$) if initially incorrect. Our metric explicitly leverages this base-rate asymmetry. As Figure 2 illustrates, unseen documents (bottom) have lower baseline accuracy, creating greater scope for in-context learning (Brown et al., 2020) to correct initially incorrect predictions (green). Conversely, memorized documents (top) possess strong exact-match priors (Carlini et al., 2023); introducing context actively interferes with these priors (Xie et al., 2024; Zawalski et al., 2026), systematically triggering the large available pool of negative shifts (red) and capturing the memorization signal without log-probabilities.

Seen Document: `pile_uspto_backgrounds` *Context hurts: 2 helped (green), 11 hurt (red)*

With a semiconductor testing system (including so-called an IC tester, LSI tester, and so forth; hereinafter referred to merely as a tester), it is necessary to supply a device as a test subject {hereinafter referred to as a DUT (Device Under Test)} with a voltage with high precision in order to inspect and test the DUT with good precision. Accordingly, a tester is provided with a power supply assembly capable of outputting a voltage with high precision. Determination on whether or not a DUT is acceptable has been made by taking measurements on

. . . . . .

Unseen Document: `patent_2025` *Context helps: 10 helped (green), 1 hurt (red)*

1. Field of the Disclosure The present disclosure generally relates to an electronic device and a method of manufacturing an electronic device. 2. Description of the Related Art Antenna package(s) having antennas for signal (e.g., radio frequency (RF) signal) transmission and reception may include an antenna layer and an RF routing layer electrically connected thereto. In some existing techniques for forming an electromagnetic interference (EMI) shielding layer on an antenna package, a tape (e.g., a thermally stable tape) may be utilized to protect the antenna

Figure 2: Positions where adding context ($K = 3$) flips the top-1 prediction (evaluated on Pythia-410M). Green highlights ($\delta_j = 1$) indicate shifts toward the ground-truth token, while red highlights ($\delta_j = -1$) indicate shifts away. Crucially, for unseen text there are significantly more tokens where the context helps (green), while for seen text the context predominantly hurts (red) by interfering with exact memorization.

2.2 Budgeted Auditing and Linear Amplification

Auditing all T positions requires $2 \cdot |\mathcal{D}| \cdot T$ queries, which is cost-prohibitive. We reduce this by evaluating a subset $\mathcal{D}_n \subseteq \mathcal{D}$ ($|\mathcal{D}_n| = n$) and limiting queries to W positions via a selection policy π_t.

Targeting the budget W exclusively on relevant tokens ($\delta_j \neq 0$) accelerates score convergence by T/T_{rel}. We formalize this via Bernstein's inequality (Boucheron et al., 2013). Let T_{rel} denote the relevant token count, defining a sparsity ratio $\rho = T_{\text{rel}}/T$. Estimating the conditional mean signal $\mu = \mathbb{E}[\delta_j \mid \delta_j \neq 0]$ (which determines the score $\mathcal{S} = (1-\mu)/2$), uniformly sampling W tokens gives expected shift $\mathbb{E}[Y_i] = \rho\mu$ and variance $\text{Var}(Y_i) \leq \rho$. For an unbiased empirical mean $\hat{\mu}_{\text{naive}} = \frac{1}{\rho W}\sum_{i=1}^{W} Y_i$, the deviation $\epsilon \in (0,1)$ is bounded by:

$$\mathbb{P}\left(|\hat{\mu}_{\text{naive}} - \mu| \geq \epsilon\right) = \mathbb{P}\left(\left|\frac{1}{W}\sum_{i=1}^{W} Y_i - \rho\mu\right| \geq \rho\epsilon\right) \leq 2\exp\left(-\frac{W\epsilon^2\rho}{2(1+2\epsilon/3)}\right) \tag{2}$$

Conversely, if an ideal targeted policy strictly selects relevant tokens, the expected shift is $\mathbb{E}[Z_i] = \mu$. The empirical mean $\hat{\mu}_{\text{target}} = \frac{1}{W}\sum_{i=1}^{W} Z_i$ bounds the error without the sparsity penalty:

$$\mathbb{P}\left(|\hat{\mu}_{\text{target}} - \mu| \geq \epsilon\right) \leq 2\exp\left(-\frac{W\epsilon^2}{2(1+2\epsilon/3)}\right) \tag{3}$$

To achieve identical statistical error bounds, the required budgets must satisfy $W_{\text{naive}}\,\rho = W_{\text{target}}$. Substituting $\rho = T_{\text{rel}}/T$, we arrive at the sample complexity:

$$W_{\text{naive}} = \frac{1}{\rho}W_{\text{target}} = \frac{T}{T_{\text{rel}}}\,W_{\text{target}} \tag{4}$$

This rate is tight, since $\text{Var}(\hat{\mu}_{\text{naive}}) = (1/\rho - \mu^2)/W \geq \text{Var}(\hat{\mu}_{\text{target}})/\rho$.

Because the membership signal is highly sparse ($T_{\text{rel}} \ll T$), blind full-sequence auditing is inefficient, and practical heuristics must approximate the idealized Oracle targeting. Crucially, memorization typically manifests in localized, continuous clusters (Carlini et al., 2023). Uniform sampling ignores this autocorrelation, wasting budget on uninformative regions.

2.3 Token and Context Selection Policies

To overcome this, the selection policy $\pi_t(x, W)$ must maximize the yield of non-zero δ_j. We evaluate:

- **Uniform / Prefix / Suffix**: Sample W positions uniformly (cf. Xu et al., 2024b), or select the first/last W positions. While Prefix often captures early shifts, Uniform and Suffix waste budget on uninformative positions ($\delta_j = 0$), incurring the full $\mathcal{O}(1/\rho)$ sparsity penalty.
- **LCS (Ours)**: Select W tokens within the Longest Common Subsequence (LCS) (Hirschberg, 1975) between target x and context C. This exploits LLM memorization structure: models regurgitate sequences in continuous blocks (Carlini et al., 2021; Nasr et al., 2025). Beyond recovering density, this introduces a beneficial selection bias: by targeting regions most likely to collide with memorized exact-match priors, LCS filters spurious contextual noise, allowing it to outperform even the idealized Oracle in high-variance regimes (Appendix C.3).
- **Oracle**: Selects strictly known $\delta_j \neq 0$ positions (idealized reference).

Context Selection (π_c). We construct context C by ranking remaining documents in $\mathcal{D}$ via Jaccard similarity (Jaccard, 1912) to x, extracting $\tau = 100$ tokens from the top $K = 3$ matches (cf. Liu et al., 2022) (ablated in Appendix C.1).

3 Budget-Accuracy Trade-Off

Experimental Setup. We evaluate the framework on 11 pre-trained off-the-shelf models (detailed in Section 4), relying strictly on base checkpoints to isolate pre-training memorization from fine-tuning variance. We compare the black-box heuristics (*Prefix, Suffix, Uniform, LCS*) against the *Oracle*, an idealized reference rather than a practical policy, since identifying $\delta_j \neq 0$ a priori requires querying every position.

Convergence of Token Selection Policies. Figure 3 validates our theoretical framework. The *Oracle* policy rapidly achieves optimal separation, as predicted in Section 2.2: only $\rho \approx 0.1$ of positions are relevant, implying a $1/\rho \approx 9.5\times$ budget reduction over *Uniform*, and we observe at least $8.8\times$ at AUC 0.75. Among black-box heuristics, *Uniform* and *Suffix* converge slowest, indicating later tokens yield minimal information gain. Conversely, *Prefix* captures relevant positions faster due to higher marginal entropy of early tokens (Olsson et al., 2022). Crucially, *LCS* is the most sample-efficient heuristic, saturating detection at $W \leq 30$ and closely tracking the Oracle. Naive policies catch up only at near-exhaustive budgets ($W \geq 100$), which negate the premise of a budgeted audit. LCS also acts as a regularizing filter, isolating verbatim copying from spurious perturbations (Appendix C.3).

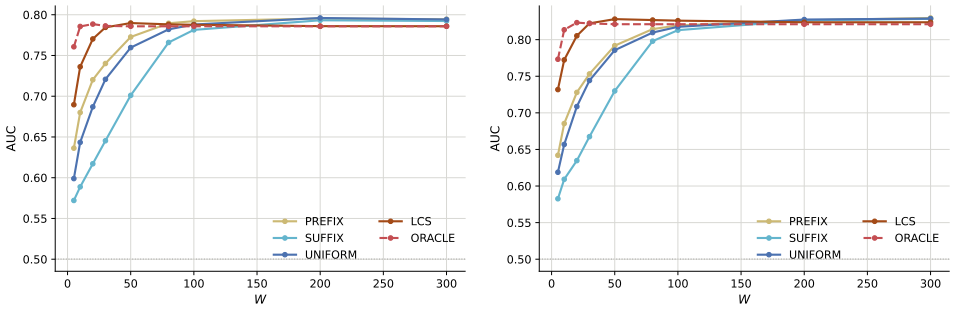

Figure 3: Cost-effectiveness of token selection policies across 15 models (11 pre-trained base models on the left and 4 post-trained Pythia (Biderman et al., 2023) checkpoints on the right). Pre-trained (left) and post-trained (right) results show mean AUC per budget. LCS demonstrates superior sample efficiency, converging to the Oracle at $W \leq 30$ and filtering spurious noise.

Hyperparameter Ablation. To optimize large-scale API audits, we grid-search over the number of context documents K and the token budget W. Figure 4 maps the mean AUC across the 7 ablation models. Performance exhibits a steep initial gain that plateaus rapidly. Utilizing $K = 3$ documents and a strictly constrained budget of $W = 16$ tokens achieves an AUC of 0.74. Expanding to $K = 5$ documents and $W \in \{64, 100\}$ tokens yields only a marginal improvement to 0.75. Consistent with Appendix C.1, the signal depends far more on *which* context documents are selected than on *how many*, so K acts primarily as a cost parameter. We therefore adopt $K = 3$ and $W = 16$ as a cost-effective configuration. This plateau corroborates the signal sparsity predicted in Section 2.2: beyond 16 selected tokens, evaluating the remaining sequence is largely redundant.

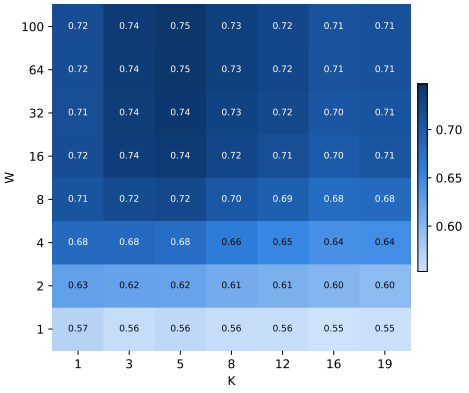

Figure 4: Hyperparameter ablation grid across 7 off-the-shelf models, displaying the AUC given varying token budgets (W) and document counts (K).

4 Auditing Public Models

Comparison with black-box methods. Following the protocol of Zawalski et al. (2026), we evaluate on 11 open-weights base checkpoints with known training data (Black et al., 2021; Biderman et al., 2023; Peng et al., 2023; Groeneveld et al., 2024), spanning the Pile (Gao et al., 2020) and Dolma (Soldaini et al., 2024) families. Our unseen dataset pools strictly post-date each family's pre-training cutoff (Appendix B); since temporal splits can introduce distribution shifts unrelated to membership (Duan et al., 2024; Das et al., 2025; Meeus et al., 2025), we discuss this limitation in Section 6.

We compare LCS against generation-based black-box baselines: verbatim extraction (Carlini et al., 2021; 2023), SA-MIA (Kaneko et al., 2024), and SimMIA (Yi & Li, 2026), the latter two under their original configurations. Since the original extraction attack requires log-probabilities, we use a graded black-box variant of the protocol of Carlini et al. (2023): from a 50-token prefix, we greedily decode 50 tokens and score the normalized length of their longest prefix matching the ground truth. We audit 100 documents per dataset with $W = 16$ and $K = 3$ (Section 3).

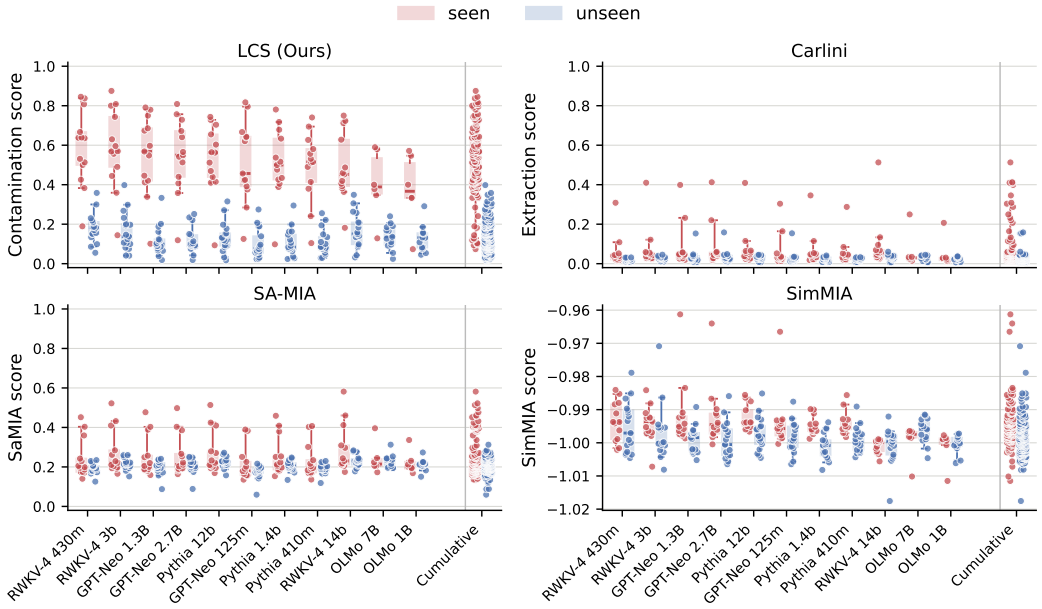

Figure 5: Contamination scores across 11 base checkpoints. Red and blue denote seen and unseen corpora, respectively. Box plots show quartiles and 5–95% whiskers with dots as individual corpora. The 2x2 grid compares our approach against standard black-box baselines. **Top-left:** Our LCS method achieves clear separation between seen and unseen data (mean AUC 0.78). **Remaining panels:** The Carlini, SA-MIA, and SimMIA baselines fail to separate distributions, collapsing to near-chance performance (mean AUCs $\approx$ 0.54–0.56).

As detailed in Table 1 and Figure 5, LCS vastly outperforms all baselines (mean AUC 0.780). Standard IID splits (e.g., train vs. validation) are ill-suited for corpus-level contamination, as members and non-members share substantial n-gram overlap, blurring generalization and memorization (Duan et al., 2024). Against domain-related, temporally separated controls, separation is striking: `pile_stackexchange` vs. `gsm8k_2021` (Cobbe et al., 2021) (AUC 0.930) and `math500_2021` (Hendrycks et al., 2021b; Lightman et al., 2024) (AUC 0.761), or `pile_github` vs. `humaneval_2021` (Chen et al., 2021) (AUC 0.896). Mechanistically, domain shift alone is difficult to reconcile with helpful semantic context actively *hurting* predictions on seen datasets ($\delta_j = -1$); this interference is instead consistent with the disruption of exact-match verbatim priors (Xie et al., 2024; Zawalski et al., 2026).

Conversely, Carlini's attack and SA-MIA collapse to near-chance performance (AUC ≈ 0.54), with seen and unseen distributions overlapping almost completely (Figure 5), while SimMIA improves only moderately. We attribute this failure to their reliance on (near-)verbatim reproduction during free-running generation—an excessively high bar for single-epoch pre-training prose (Duan et al., 2024). These baselines only registered elevated scores on templated, low-entropy corpora (e.g., `pile_github`), consistent with extraction being dominated by duplicated sequences (Kandpal et al., 2022; Carlini et al., 2023).

Table 1: Black-box baseline comparison, per model. AUC separating each model's seen corpora from its unseen ones, pooled over documents.

Model	Carlini	SA-MIA	SimMIA	LCS (Ours)
GPT-Neo 125M	0.526	0.578	0.577	**0.822**
GPT-Neo 1.3B	0.536	0.578	0.609	**0.808**
GPT-Neo 2.7B	0.540	0.565	0.606	**0.812**
Pythia 410M	0.557	0.565	0.607	**0.811**
Pythia 1.4B	0.562	0.569	0.613	**0.816**
Pythia 12B	0.558	0.559	0.583	**0.796**
RWKV-4 430M	0.559	0.572	0.544	**0.785**
RWKV-4 3B	0.562	0.565	0.600	**0.792**
RWKV-4 14B	0.572	0.557	0.474	**0.724**
OLMo 1B	0.487	0.459	0.491	**0.696**
OLMo 7B	0.504	0.452	0.435	**0.717**
Mean	0.542	0.547	0.558	**0.780**

Unlike instance-level MIAs (Shokri et al., 2017), which perform near chance on real pre-training data due to near single-epoch training and fuzzy membership boundaries (Duan et al., 2024), dataset-level approaches aggregate weak per-instance signals (Maini et al., 2024; Puerto et al., 2025; Zawalski et al., 2026). Our approach extends this idea to strictly black-box settings by aggregating targeted, context-induced prediction shifts across an entire corpus, without requiring unprompted verbatim reproduction.

Table 2: Comparison of our black-box method (LCS) to competing gray-box methods requiring log-probabilities.

Method	AUC	TPR@5%FPR
Loss	0.589	0.129
Min-K%	0.580	0.118
zlib	0.771	0.279
CoDeC	**0.856**	**0.350**
LCS (Ours)	0.780	0.173

Comparison with gray-box methods. Despite the query constraint ($W = 16$), our black-box LCS policy (AUC 0.780; Table 2) outperforms Loss (Yeom et al., 2018) (0.589) and Min-K% (Shi et al., 2024) (0.580) and matches zlib (Carlini et al., 2021) in AUC (0.771), although zlib is stronger at low false-positive rates (TPR@5%FPR 0.279 vs. 0.173). Uncalibrated likelihood scores such as Loss and Min-K% conflate memorization with the intrinsic predictability of text, which zlib partially corrects by normalizing against compression entropy (Carlini et al., 2021); our metric plays an analogous calibrating role, as each position is compared against its own unconditioned prediction. The gap to CoDeC (Zawalski et al., 2026) reflects the cost of observing binary argmax shifts instead of log-probabilities; still, LCS recovers roughly 80% of CoDeC's above-chance AUC.

Table 3: Detection performance on post-trained checkpoints. The contamination signal is largely preserved across SFT and alignment.

Model	AUC	TPR@5%FPR
Pythia 410M	0.811	0.289
+ SFT	0.805	0.143
+ DPO	0.801	0.148
Pythia 1.4B	0.816	0.158
+ SFT	0.803	0.240
+ DPO	0.797	0.132
RWKV-4 3B	0.792	0.214
+ SFT	0.722	0.079
RWKV-4 14B	0.724	0.051
+ SFT	0.725	0.051
OLMo 7B	0.717	0.085
+ SFT	0.688	0.110
+ Instruct	0.700	0.117

Audit of post-trained models. We evaluate aligned checkpoints (Peng et al., 2023; Groeneveld et al., 2024) to test whether post-training obscures the membership signal (Table 3). While SFT and DPO (Ouyang et al., 2022; Rafailov et al., 2023) suppress unprompted regurgitation (Nasr et al., 2025), our ground-truth prefixes avoid free-running rollouts altogether. AUC drops by at most 0.03 for most checkpoints, with the largest degradation for RWKV-4 3B after SFT (0.792 to 0.722), suggesting that alignment largely masks, rather than erases, pre-training memorization (Nasr et al., 2025; Lin et al., 2024; Qi et al., 2025).

5 Auditing Frontier Models

To demonstrate the practical feasibility of our framework, we deploy it against frontier models via public inference APIs under strict black-box constraints. Lacking public ground truth for the training mixtures of these commercial systems, this audit serves as an exploratory proof-of-concept rather than a definitive claim of contamination (Zhang et al., 2025a). To query near-deterministic next-token predictions, we dispatch requests at temperature 0 (0.6 for Kimi-k2.6, due to provider constraints), allocate token positions via LCS, and exclusively use endpoints supporting assistant prefill to isolate the memorization signal from instruction bias. Prefill preserves plain continuation: the target token is the top-1 prediction at 72.6% of positions with prefill versus 74.1% without any template (Appendix C.4), so the API audits probe the same next-token behavior as our open-weights experiments.

We evaluate Kimi-k2.6 (Kimi Team, 2026; Moonshot AI, 2026), DeepSeek-V3 (DeepSeek-AI, 2025), DeepSeek-V4-Pro (DeepSeek-AI, 2026), Claude Sonnet 4.5 (Anthropic, 2025b), and Claude Opus 4.5 (Anthropic, 2025a). Allocating a budget of $K = 3$ context documents and $W = 16$ tokens for $n = 100$ sequences yields a highly efficient upper bound of $B \leq 3{,}200$ API calls per evaluation. A detailed breakdown is provided in Appendix B.3.

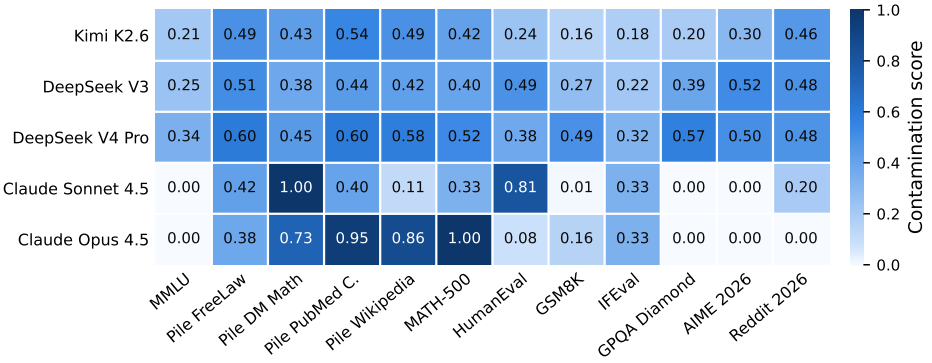

Figure 6: Contamination scores ($\mathcal{S}$) for frontier models evaluated via public APIs. Recent datasets (`aime2026`, `reddit_2026`) act as internal negative controls to account for varying base capabilities, exposing pronounced behavioral gaps on established pre-training corpora.

As base capabilities and alignment strictness vary across models, we replace a universal threshold with a per-model baseline built from post-2025 datasets. These controls are smaller (e.g., $n = 25$ for `aime2026`), so the baseline exhibits higher variance. Moreover, since the training cutoffs of the most recent models (Kimi-k2.6, DeepSeek-V4-Pro) are not publicly specified, we cannot rule out partial exposure to these controls, which would bias their baselines upward and make our estimates conservative.

Three of the five models separate pre-cutoff data from their 2026 controls (Figure 6). Claude Opus 4.5 scores near the maximum on MATH-500 (Hendrycks et al., 2021b; Lightman et al., 2024), `pile_pubmed_central`, and `pile_wikipedia`, yet zero on both controls, including the similarly formatted math control `aime2026`. Claude Sonnet 4.5 shows the same pattern on `pile_dm_mathematics` and HumanEval (Chen et al., 2021), and DeepSeek-V4-Pro, less sharply, on `pile_freelaw`, `pile_pubmed_central`, and GPQA Diamond (Rein et al., 2024).

Figure 6 also reveals the difficulty of auditing frontier models. Despite similar controls, Opus and Sonnet disagree anticipated positive controls: `pile_wikipedia` and HumanEval. Kimi-k2.6 (which we could only query at non-zero temperature) and DeepSeek-V3 do not separate, with expected positives staying close to their 2026 controls. MMLU (Hendrycks et al., 2021a) scores at or below the controls on all five models. We attribute the weak signal on saturated text to the base-rate mechanics of Section 2.1. As frontier models approach near-perfect unconditioned accuracy on ubiquitous prose or structured benchmarks such as MMLU (cf. Carlini et al., 2023), the pool of context-induced flips ($\delta_j \neq 0$) shrinks, and the score of such a dataset rests on few relevant positions. Argmax-based audits therefore lose sensitivity as model capabilities saturate, motivating dynamic difficulty scaling in future audits (Zhu et al., 2024; White et al., 2025).

6 Discussion

Observing only discrete output changes under small query budgets suffices to rival several full-access methods, lowering the barrier to third-party verification of copyright compliance and benchmark integrity on commercial APIs (Henderson et al., 2023; Sainz et al., 2023). However, black-box evidence alone cannot prove that specific data was used in training (Casper et al., 2024; Zhang et al., 2025a).

Limitations and Future Work. Our framework faces four main constraints. First, it relies on assistant prefill, which some providers have already withdrawn for newer models, motivating adaptations to strictly conversational APIs. Second, deliberate provider obfuscation (e.g., output perturbations) could disrupt the metric (Carlini et al., 2024). Third, our negative controls rely on temporally separated corpora, which may introduce distribution shifts unrelated to membership (Duan et al., 2024; Das et al., 2025; Meeus et al., 2025); model-free blind baselines would complement our analysis. Finally, as frontier models grow more capable, their unprompted predictions naturally align closer with ground truth text, compressing the observable signal (cf. Carlini et al., 2023). Future research should explore dynamic difficulty scaling (Zhu et al., 2024; White et al., 2025), e.g., targeting complex or low-frequency text regions (cf. Shi et al., 2024; Zhang et al., 2025b).

7 Related Work

In-Context Conditioning and Memorization. LLMs memorize training data (Tirumala et al., 2022; Carlini et al., 2023), while in-context copying mechanisms such as induction heads (Olsson et al., 2022) let the prompt override these memorized priors. Context-induced shifts can thus quantify contamination (Xie et al., 2024; Zawalski et al., 2026; Yi & Li, 2026); we formalize this as information gain (Appendix A) and show that a discrete indicator retains most of this signal.

Gray-Box Data Contamination. Gray-box detectors score model likelihoods (Yeom et al., 2018) with zlib normalization (Carlini et al., 2021), focus on low-probability tail tokens (Shi et al., 2024; Zhang et al., 2025b), or calibrate via perturbations (Mattern et al., 2023; Mitchell et al., 2023), token frequencies (Zhang et al., 2024), or conditional prefixes (Xie et al., 2024). Others test benchmark-order exchangeability (Oren et al., 2024), learn detectors over output distributions (Bar-Shalom et al., 2026), or aggregate instance scores into dataset-level tests (Maini et al., 2024; Puerto et al., 2025). All require log-probabilities, which major providers increasingly restrict or disable (Carlini et al., 2024; Finlayson et al., 2024).

Black-Box Contamination Detection. For sampling-only APIs (Casper et al., 2024), instruction-dependent methods force structured tasks (Golchin & Surdeanu, 2024; Deng et al., 2024; Duarte et al., 2024), conflating instruction failure with non-memorization. Generation-based alternatives rely on verbatim extraction (Carlini et al., 2021; Nasr et al., 2025), n-gram prediction accuracy (Xu et al., 2024b), output-distribution peakedness (Dong et al., 2024), or pseudo-likelihoods (Kaneko et al., 2024; Yi & Li, 2026). These require either repeated free-running generations, prone to error propagation (Bengio et al., 2015), or untargeted position sampling, incurring high query costs. We instead combine targeted token selection with teacher forcing (Williams & Zipser, 1989) to isolate informative "tells."

8 Conclusion

In this work, we bridge the gap between gray-box contamination metrics and restrictive commercial APIs. By targeting the sparse membership signal with heuristics like LCS, we drastically reduce query budgets relative to full-sequence black-box audits, while our contamination score rivals several full-access methods. This provides a practical foundation for rigorous, independent audits of models in the wild.

AI Use Statement

In this work, we used generative AI tools to draft sections of the paper, aid and polish the writing of the manuscript, and as a coding assistant to help implement and debug our experimental pipeline. We have not used generative AI tools to formulate the core scientific claims or to generate the empirical results presented in this study.

We have reviewed all AI-assisted work. Specifically, all AI-generated code was systematically verified and tested for correctness by the authors to ensure accurate execution of the API evaluations and metric calculations. Additionally, all AI-drafted and polished text was rigorously reviewed, edited, and validated by the authors to guarantee scientific accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement

This research aims to promote transparency and accountability in Large Language Models (LLMs) by enabling independent, budget-constrained audits for data contamination. While Membership Inference Attacks can raise data privacy concerns, our methodology is explicitly designed for corpus-level detection rather than instance-level extraction. Crucially, the framework requires the auditor to already possess the target document to construct the conditional prefix, meaning it cannot be used by malicious actors to blindly extract novel Personal Identifiable Information (PII) or sensitive intellectual property.

Additionally, all evaluations on commercial models were conducted via public APIs in strict compliance with standard usage terms. Through mathematical optimization, our approach drastically minimizes the compute overhead and energy consumption typically associated with generation-based audits, ensuring that large-scale transparency research can be performed sustainably.

Reproducibility Statement

To ensure full reproducibility of our results, we have made the complete source code publicly available at `https://anonymous.4open.science/r/blackbox-contamination-detector-C4CD/README.md`. This repository includes the reference implementation of our black-box estimator as well as a standalone tool for conducting custom dataset audits. All experimental setups, including data partitioning, model checkpoints, and LoRA fine-tuning hyperparameters, are exhaustively documented in Appendix B and C.3. Finally, the exact prompt structures and provider-specific configurations necessary to replicate the frontier model API evaluations are outlined in Appendix B.3.

References

Anthropic. System card: Claude opus 4.5. https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47.pdf, 2025a.

Anthropic. System card: Claude sonnet 4.5. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd/Claude%20Sonnet%204.5%20System%20Card.pdf, 2025b.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Guy Bar-Shalom, Fabrizio Frasca, Derek Lim, Yoav Gelberg, Yftah Ziser, Ran El-Yaniv, Gal Chechik, and Haggai Maron. Beyond next token probabilities: learnable, fast detection of hallucinations and data contamination on llm output distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i36.40254. URL https://doi.org/10.1609/aaai.v40i36.40254.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/e995f98d56967d946471af29d7bf99f1-Paper.pdf.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX*

Security Symposium (USENIX Security 21), pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=VE3yWXt3KB.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2254–2272, 06 2024. doi: 10.1145/3630106.3659037.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://arxiv.org/abs/2501.14249.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 12 1934. ISSN 0006-3444. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind Baselines Beat Membership Inference Attacks for Foundation Models . In *2025 IEEE Security and Privacy Workshops (SPW)*, pp. 118–125, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/SPW67851.2025.00016. URL https://doi.ieeecomputersociety.org/10.1109/SPW67851.2025.00016.

DeepSeek-AI. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716. URL https://aclanthology.org/2024.findings-acl.716/.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. De-cop: detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, January 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRcYFm8vyB.

Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen, and Fei Xia. Does data contamination detection work (well) for LLMs? a survey and evaluation on detection assumptions. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5250–5271, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.291. URL https://aclanthology.org/2025.findings-naacl.291/.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24 (400):1–79, 2023. URL http://jmlr.org/papers/v24/23-0569.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, 18(6):341–343, June 1975. ISSN 0001-0782. doi: 10.1145/360825.360861. URL https://doi.org/10.1145/360825.360861.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.147. URL https://aclanthology.org/2026.acl-long.147/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2): 37–50, 1912. doi: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x. URL https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models, 2022. URL https://arxiv.org/abs/2202.06539.

Masahiro Kaneko, Youmi Ma, Yuki Wata, and Naoaki Okazaki. Sampling-based pseudo-likelihood for membership inference attacks, 2024. URL https://arxiv.org/abs/2404.11262.

Kimi Team. Kimi k2: Open agentic intelligence, 2026. URL https://arxiv.org/abs/2507.20534.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.),

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wxJ0eXwwda.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (eds.), *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.18. URL https://aclanthology.org/2022.acl-short.18/.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 124069–124092. Curran Associates, Inc., 2024. doi: 10.52202/079017-3941. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e01519b47118e2f51aa643151350c905-Paper-Conference.pdf.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.719. URL https://aclanthology.org/2023.findings-acl.719/.

Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. SoK: Membership Inference Attacks on LLMs are Rushing Nowhere (and How to Fix It) . In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 385–401, Los Alamitos, CA, USA, April 2025. IEEE Computer Society. doi: 10.1109/SaTML64287.2025.00028. URL https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00028.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Moonshot AI. Kimi K2.6 tech blog: Advancing open-source coding. https://www.kimi.com/blog/kimi-k2-6, 2026.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vjel3nWP2a.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, September 2022.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936/.

Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.234. URL https://aclanthology.org/2025.findings-naacl.234/.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. A comprehensive survey of contamination detection methods in large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=SxNMjbtdFm.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models . In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, May 2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=u3vEuRr08MT.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal,

Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280, 1989. URL http://dblp.uni-trier.de/db/journals/neco/neco1.html#WilliamsZ89.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. ReCaLL: Membership inference via relative conditional log-likelihoods. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.493. URL https://aclanthology.org/2024.emnlp-main.493/.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024a. URL https://arxiv.org/abs/2406.04244.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models, 2024b. URL https://arxiv.org/abs/2404.18824.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023. URL https://arxiv.org/abs/2311.04850.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Jiatong Yi and Yanyang Li. Membership inference on llms in the wild, 2026. URL https://arxiv.org/abs/2601.11314.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in LLMs via in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YlpaaYxx4t.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramer. Position: Membership Inference Attacks Cannot Prove That a Model was Trained on Your Data . In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 333–345, Los Alamitos, CA, USA, April 2025a. IEEE Computer Society. doi: 10.1109/SaTML64287.2025.00025. URL https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00025.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=ZGkfoufDaU.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Pretraining data detection for large language models: A divergence-based calibration method. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.300. URL https://aclanthology.org/2024.emnlp-main.300/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjfOL9z5Xr.

A Theoretical Foundations

Let θ mean parameters of the model that has seen dataset $\mathcal{D}$ during training and ϕ denote the uncontaminated parameters. Information-theoretical perspective on various data contamination/data inference scoring among many different types of attacks (Carlini et al., 2021; Mitchell et al., 2023; Mattern et al., 2023; Xie et al., 2024; Zawalski et al., 2026) reveals the underlying use of the difference of conditioned and marginal distribution entropy shift i.e.

$$H(x) - H(x \mid C) = \mathrm{IG}(x, C) > 0 \tag{5}$$

thus, the effect is amplified with respect to what context C was used and how strongly it narrows down the entropy of the model distribution. When a model has seen $\mathcal{D}$ during training, much of C is already encoded in θ, therefore, $\mathrm{IG}_\theta(x, C) \to 0$ (Zawalski et al., 2026). Conversely, $\mathrm{IG}_\phi(x, C) \gg 0$ as $C \subseteq \mathcal{D}$ is a highly-informative descriptor of the audited dataset. Naturally, $\Delta_{\log} \approx \mathrm{IG}(x, C)$. Δ_{bb} is a projection of the logprob shift onto a comparison of two Kronecker deltas, equivalently approximating, albeit more crudely, the context-induced information gain.

Remark on Teacher Forcing. The statistical guarantees provided in Section 2.2 rely on the assumption that the sampled indicator shifts Y_i are drawn independently. While standard autoregressive modeling violates this, our framework operates strictly in a teacher-forcing regime (Williams & Zipser, 1989). For every evaluated position j, the prompt is constructed using the true ground-truth prefix $x_{<j}$ from the target dataset, rather than the model's own prior generations. Because of this, errors cannot propagate between positions (Bengio et al., 2015), and each δ_j is a fixed function of the document that does not depend on which other positions are queried; positions drawn at random therefore yield independent shifts Y_i, satisfying the independence assumption (see Hoeffding, 1963, Section 6).

Autocorrelation of the True Signal. While the sampling process is independent due to teacher forcing, the underlying true membership signal within the text is strongly autocorrelated. Memorization occurs in continuous chunks rather than uniformly distributed single tokens (Carlini et al., 2023; Nasr et al., 2025). Uniform sampling fails to exploit this structure, thereby paying the full $\mathcal{O}(\frac{1}{\rho})$ sparsity penalty derived in our budget allocation proof. Conversely, policies like LCS explicitly exploit this autocorrelation by targeting dense regions of lexical overlap, bridging the gap to the theoretical optimum, although positions within one window are then correlated, so the guarantees of Section 2.2 apply to them only approximately.

B Detailed Experimental Setup

B.1 Public Models

We evaluate 11 base checkpoints spanning diverse parameter scales (125M to 14B) whose pre-training corpora (The Pile and Dolma) (Gao et al., 2020; Soldaini et al., 2024) are publicly accessible, so the membership of the seen corpora is known. Table 4 lists the base models and post-trained derivatives, including RWKV-4 as a non-transformer control architecture.

B.2 Datasets

For each model family, we pair *seen* subsets of its pre-training mixture with *unseen* negative-control datasets.

Seen Datasets (The Pile): `github`, `wikipedia`, `stackexchange`, `pubmed_abstracts`, `uspto_backgrounds`, `cc`, `openwebtext2`, `pubmed_central`, `freelaw`, `hackernews`, `nih_exporter`, `arxiv`, `dm_mathematics` (Gao et al., 2020).

Seen Datasets (Dolma): `c4`, `wiki`, `pes2o`, `reddit`, `starcoder`, `cc` (Soldaini et al., 2024).

Table 4: Model families used in Sections 3 and 4 (Biderman et al., 2023; Black et al., 2021; Peng et al., 2023; Groeneveld et al., 2024).

Family	Base Sizes	SFT	RL / Aligned
Corpus: The Pile			
Pythia	410M, 1.4B, 12B	410M, 1.4B	410M, 1.4B[†]
GPT-Neo	125M, 1.3B, 2.7B	–	–
RWKV-4	430M, 3B, 14B	3B, 14B[‡]	–
Corpus: Dolma			
OLMo	1B, 7B	7B	7B

[†]Includes DPO (Rafailov et al., 2023) (410M, 1.4B) and PPO (Schulman et al., 2017) (1.4B). [‡]Raven checkpoints.

Unseen Datasets (Shared): `arxiv_2025`, `bbc_news_2025`, `patent_2025`, `gutenberg_fiction`, `aime2025`, `aime2024`, `bbq` (Parrish et al., 2022), `reward_bench_v1` (Lambert et al., 2025), `frames` (Krishna et al., 2025), `gpqa_diamond` (Rein et al., 2024), `amazon_reviews_2023` (Hou et al., 2026), `global_news`, `livecodebench_v1/v5` (Jain et al., 2025). (*Note:* `gsm8k_2021` (Cobbe et al., 2021), `math500_2021` (Hendrycks et al., 2021b; Lightman et al., 2024), `humaneval_2021` (Chen et al., 2021), and `ifeval_2023` (Zhou et al., 2023) were additionally used as unseen controls exclusively for Pile-trained models).

B.3 Frontier Models and API Logistics

We evaluate Claude Sonnet and Opus 4.5 (Anthropic, 2025b;a) via OpenRouter, DeepSeek-V3 and V4-Pro (DeepSeek-AI, 2025; 2026) through identical API routing, and Kimi-k2.6 (Kimi Team, 2026; Moonshot AI, 2026) directly via Moonshot AI.

To isolate pure memorization from instruction-following bias, all requests use assistant prefill (M-mode; Appendix C.4), forcing direct sequence continuation. For DeepSeek-V4-Pro, reasoning tokens are disabled. Rate limits and transient API failures are handled by an asynchronous execution wrapper with exponential backoff, preventing dropped queries.

Generation parameters are standardized with a 512-token cap and temperature 0 to minimize sampling stochasticity, except for Kimi-k2.6, which uses temperature 0.6 due to provider constraints. We target $n = 100$ documents per dataset, with minor deviations: `pile_dm_mathematics` contains 99 valid documents, `aime2026` yields 25 after filtering items under 200 characters, and `reddit_2026` provides 99 valid matches.

Using $K = 3$ context documents, the theoretical ceiling is $W = 16$ token positions ($\leq$ 3,200 calls per dataset), though realized volume scales down with length and overlap. Unconditioned completions regularly hit the 512-token cap, driving the completion volume. Table 5 summarizes token usage and cost across the evaluation suite.

Table 5: Total API usage and costs aggregated across all dataset evaluations.

Model	Prompt Tokens	Completion Tokens	Cost (USD)
Claude Opus 4.5	8,314,517	1,280,741	$73.59
Claude Sonnet 4.5	8,314,517	853,024	$37.74
DeepSeek-V4-Pro	7,245,621	4,609,999	$9.74
DeepSeek-V3	7,200,410	857,710	$3.07
Kimi-k2.6	7,495,784	862,260	$\leq$ $8.92
Total			$\leq$ **$133.06**

C Additional Experiments

C.1 Context Construction Policy

To evaluate the robustness of our context construction approach, we ablate both the selection strategy (π_c) and the number of context documents (K). We measure performance using the separation gap $\Delta\mathcal{S} = \mathcal{S}_{\text{seen}} - \mathcal{S}_{\text{unseen}}$ and the detection AUC, averaged across two models (`gpt-neo-125m` and `pythia-410m`) under a strictly constrained token budget of $W = 16$.

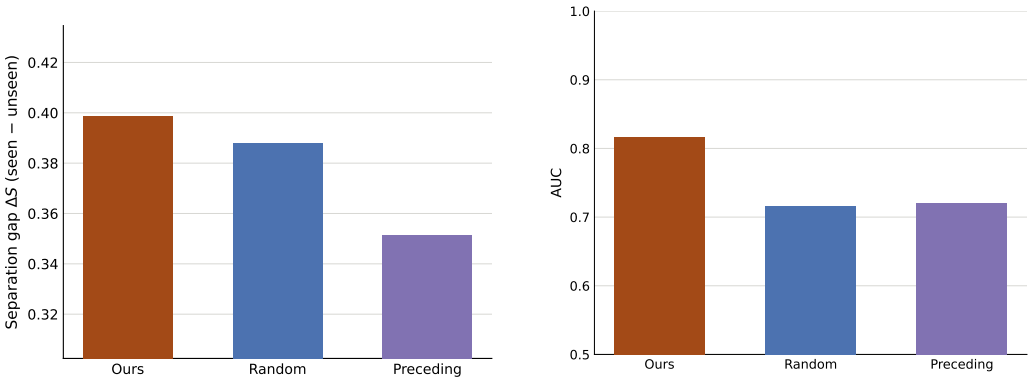

Figure 7: Context selection policy ablation. **Left:** Mean separation gap ($\Delta\mathcal{S}$). **Right:** Contamination-detection AUC. Both metrics use $K = 3$ context documents, $W = 16$ selected tokens, and are averaged over `gpt-neo-125m` and `pythia-410m`.

Context Selection Strategy. We compare three methods for constructing the prepended context C: our default *similar-prefixes* policy (Jaccard-ranked) (Jaccard, 1912), *random prefixes* (cf. Zawalski et al., 2026), and *preceding* (sequential text). As shown in Figure 7, our Jaccard-based selection provides a large advantage over both naive baselines. It yields the highest mean separation gap ($\Delta\mathcal{S} \approx 0.399$) and substantially boosts the detection performance to an AUC of 0.817. In contrast, random and sequential-preceding selections perform nearly identically in terms of AUC (0.716 and 0.721, respectively), a drop of roughly 0.10 AUC ($\approx 14\%$ relative) compared to our method. This confirms that dynamically maximizing lexical overlap is crucial for recovering a strong contamination signal (cf. Liu et al., 2022).

Context Size (K). Having established that the choice of documents is critical, we evaluate the sensitivity of the signal to the volume of context. We fix the selection policy to *similar-prefixes* and vary the number of supplied context documents K from 1 to 16. Figure 8 demonstrates that the separation gap remains highly stable, fluctuating within a narrow band of 0.392 to 0.413. Providing a single context document ($K = 1$, $\Delta\mathcal{S} = 0.401$) recovers essentially the exact same membership signal as providing sixteen documents ($K = 16$, $\Delta\mathcal{S} = 0.407$). This indicates that the separation power is highly sensitive to *which* documents are selected, but largely insensitive to *how many* are supplied. Thus, K acts mainly as a compute and cost parameter rather than an accuracy bottleneck.

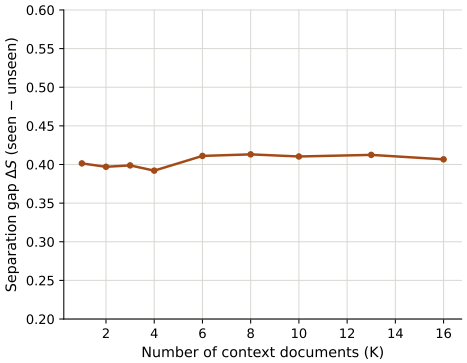

Figure 8: The effect of varying the number of context documents (K) on the separation gap.

C.2 Non-Zero Temperature Sampling

To evaluate the robustness of our framework when access to deterministic decoding is restricted, we sweep the sampling temperature from 0 to 1. We hold the audit configuration fixed at our optimal parameters ($K = 3$, $W = 16$, LCS policy) and average the results across six models (Pythia 12B, 1.4B, 410M, and GPT-Neo 2.7B, 1.3B, 125M) (Biderman et al., 2023; Black et al., 2021).

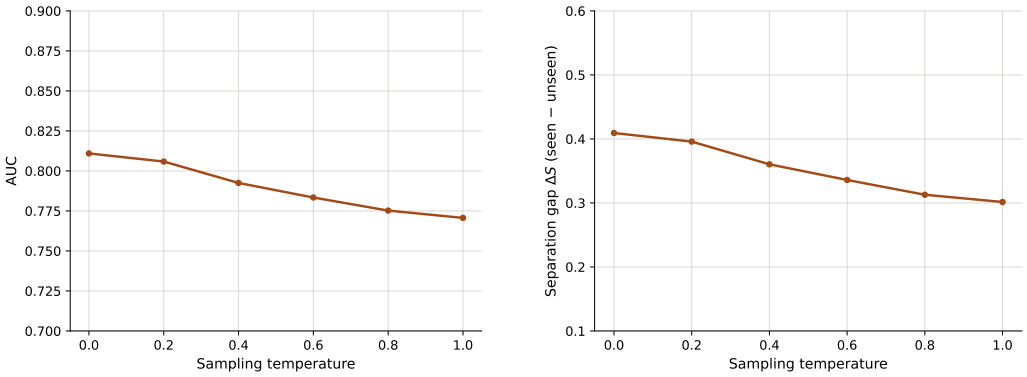

Figure 9: Detection robustness under non-zero temperature sampling. **Left:** Contamination-detection AUC as temperature varies. **Right:** Mean score separation gap ($\Delta\mathcal{S}$) between contaminated and held-out corpora. Results are averaged over 6 models.

As shown in Figure 9, contamination detection degrades gracefully, rather than catastrophically, as the sampling temperature increases. The detection AUC drops by only 5% relative, from 0.811 at temperature 0 to 0.771 at temperature 1. Concurrently, the raw separation gap ($\Delta\mathcal{S}$) shrinks more noticeably, decreasing by roughly 26% relative (from 0.409 to 0.301) over the same temperature range. These results demonstrate that while stochastic sampling compresses the margin between seen and unseen distributions, the method maintains strong discriminative power (AUC $\gg 0.5$) even under fairly aggressive temperature settings.

C.3 Robustness to High-Variance Regimes

To analyze the robustness of our token selection policies under high-variance conditions, we evaluated a controlled setup using single-epoch fine-tuning on small data subsets. Building upon the architectures evaluated in Appendix B.1, we selected an expanded set of 14 Pile-pretrained models: Pythia (160M, 410M, 1.4B, 2.8B, 6.9B, 12B) (Biderman et al., 2023), GPT-Neo (125M, 1.3B, 2.7B) (Black et al., 2021), GPT-J (6B), and RWKV-4 (430M, 3B, 7B, 14B) (Peng et al., 2023). Additionally, we constructed a contamination pool of 16 corpora released after the Pile knowledge cutoff (Gao et al., 2020). While largely overlapping with the negative controls established in Appendix B.2, this pool was specifically tailored for this experiment: arXiv abstracts, BBC news, USPTO patents, Project Gutenberg fiction, GSM8K, MATH-500, HumanEval, IFEval, BBQ, RewardBench, FRAMES, Amazon reviews, global news, LiveCodeBench, HH-RLHF-helpful (Bai et al., 2022), and TinyStories (Eldan & Li, 2023). Each corpus was capped at 600 documents.

We drew 4 random seen/unseen partitions of this pool (7–10 seen corpora each) to contaminate the models via single-epoch fine-tuning. Models $\geq$ 2B parameters used merged LoRA adapters ($r = 16$) (Hu et al., 2022), while smaller models were fully fine-tuned, yielding 56 unique checkpoints. Crucially, this unconstrained random partitioning creates asymmetric dataset exposure. Rather than enforcing an artificially balanced curriculum, this design deliberately simulates the heterogeneous and unstructured contamination typical of real-world deployments. This stress-tests the estimator's robustness to dataset-specific idiosyncrasies, ensuring the metric captures genuine memorization rather than artifacts of a perfectly symmetric evaluation.

We audited each checkpoint on all 16 corpora ($n = 100$ documents per corpus, $K = 3$ context documents) under 5 window-selection policies and 9 window budgets $W \in \{5, \ldots, 300\}$, totaling 2,520 audits.

In this setup, the single-epoch fine-tuning induces significant distributional polarization and weight shifts. The addition of the context condition C causes the models to alter their predictions sharply across the sequence, generating a high volume of output shifts ($\delta_j \neq 0$) driven by local overfitting rather than stable pre-training memorization. Consequently, the Oracle policy, which strictly aggregates all $\delta_j \neq 0$ positions, suffers a degradation in AUC due to the influx of these false-positive contextual perturbations.

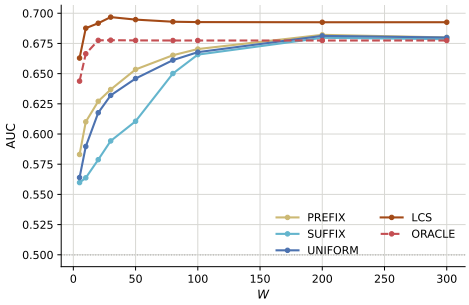

Figure 10: AUC performance of selection policies on unstable, single-epoch fine-tuned models across varying token budgets. The LCS policy filters context-induced noise and strictly outperforms the indiscriminate Oracle aggregation.

Conversely, the LCS policy strictly outperforms the Oracle policy across all budgets W in this high-variance environment. By constraining the evaluation budget to regions with high lexical overlap, LCS actively filters out the spurious contextual noise. This confirms that tying the query budget to lexical similarity metrics provides a robust regularization mechanism, ensuring the estimator reliably tracks the true membership signal even in severely overfitted, volatile, or rapidly adapted models.

C.4 M-Mode vs I-Mode Prompting

We evaluate audits that wrap each query in a chat template (Ouyang et al., 2022), which we call I-mode:

$$i_C^j = \mathbf{1}\big[\hat{x}_j = x_j \mid \texttt{<user>}\, C, x_{<j}\, \texttt{</user>}\, \texttt{<assistant>}\big], \tag{6}$$

yielding the indicator $\delta_j = i_C^j - i_\emptyset^j$. We also test assistant prefill (Qi et al., 2025), which we call M-mode:

$$i_C^j = \mathbf{1}\big[\hat{x}_j = x_j \mid \texttt{<assistant>}\, C, x_{<j}\big], \tag{7}$$

and compare both with prompting without any template (U-mode), as used in Sections 3 and 4.

Table 6: Fraction of target tokens ranked within the top-K predictions under U-mode, M-mode, and I-mode. The first row corresponds to the audit's rule.

	Target token within top-K			Change from U-mode	
Top-K	U-mode	M-mode	I-mode	ΔM	ΔI
1	0.741	0.726	0.016	−0.015	**−0.724**
2	0.805	0.796	0.018	−0.008	−0.786
3	0.836	0.833	0.022	−0.003	−0.813
5	0.859	0.857	0.028	−0.002	−0.831
10	0.886	0.886	0.086	+0.000	−0.800
20	0.914	0.911	0.133	−0.003	−0.781
50	0.950	0.948	0.197	−0.003	−0.754
100	0.965	0.964	0.230	−0.001	−0.735
500	0.988	0.990	0.298	+0.002	−0.690
1000	0.995	0.995	0.349	+0.000	−0.645
10000	0.999	0.999	0.715	+0.000	−0.285

For five randomly sampled datasets, Table 6 shows that U-mode and M-mode place the target token at top-1 in about 73% of cases, whereas I-mode does so in under 2% and shifts the entire distribution.

This failure stems from instruction framing inducing distinct operational modes (Lin et al., 2024). With prompt template χ, a language model can be viewed as the marginalization

$$p(x_t \mid x_{<t}, \chi) = \sum_z p_z(x_t \mid x_{<t})\, p(z \mid \chi) \tag{8}$$

over modes z (cf. Xie et al., 2022). In I-mode, p_{I} dominates, so $D_{KL}(p_{\mathrm{M}} \parallel p_{\mathrm{I}})$ is high, whereas $D_{KL}(p_{\mathrm{M}} \parallel p_{\mathrm{U}})$ remains low as both model standard language. Where prefill is supported, M-mode thus faithfully proxies raw continuation, consistent with alignment leaving base-model behavior largely intact (Nasr et al., 2025; Lin et al., 2024).

D Extended Audit Results

This section extends Section 4 with per-model AUC (Table 7) and TPR@5%FPR (Table 8), including 95% CIs, for Carlini (Carlini et al., 2023), SA-MIA (Kaneko et al., 2024), SimMIA (Yi & Li, 2026), and LCS. LCS outperforms all baselines in AUC on every model. SA-MIA occasionally attains higher TPR@5%FPR, likely by capturing low-entropy documents, but its near-chance AUC limits its overall reliability.

Table 7: AUC scores (95% CIs) separating seen and unseen corpora per model and method.

Model	Carlini	SA-MIA	SimMIA	LCS (Ours)
GPT-Neo 125M	0.526 [0.504,0.546]	0.578 [0.553,0.601]	0.577 [0.556,0.598]	**0.822** [0.807,0.838]
GPT-Neo 1.3B	0.536 [0.514,0.558]	0.578 [0.553,0.600]	0.609 [0.587,0.631]	**0.808** [0.792,0.821]
GPT-Neo 2.7B	0.540 [0.516,0.562]	0.565 [0.542,0.588]	0.606 [0.585,0.629]	**0.812** [0.798,0.825]
Pythia 410M	0.557 [0.533,0.580]	0.565 [0.541,0.588]	0.607 [0.585,0.630]	**0.811** [0.799,0.826]
Pythia 1.4B	0.562 [0.541,0.583]	0.569 [0.546,0.591]	0.613 [0.592,0.635]	**0.816** [0.801,0.829]
Pythia 12B	0.558 [0.536,0.582]	0.559 [0.535,0.582]	0.583 [0.561,0.604]	**0.796** [0.780,0.810]
RWKV-4 430M	0.559 [0.537,0.580]	0.572 [0.549,0.595]	0.544 [0.502,0.586]	**0.785** [0.770,0.799]
RWKV-4 3B	0.562 [0.537,0.584]	0.565 [0.543,0.589]	0.600 [0.558,0.642]	**0.792** [0.777,0.808]
RWKV-4 14B	0.572 [0.550,0.596]	0.557 [0.533,0.581]	0.474 [0.429,0.516]	**0.724** [0.709,0.741]
OLMo 1B	0.487 [0.459,0.517]	0.459 [0.427,0.491]	0.491 [0.463,0.519]	**0.696** [0.671,0.721]
OLMo 7B	0.504 [0.473,0.535]	0.452 [0.419,0.484]	0.435 [0.406,0.463]	**0.717** [0.696,0.743]
Mean	0.542	0.547	0.558	**0.780**

Table 8: TPR@5%FPR (95% CIs) across models. Baselines use percentile bootstraps (Efron, 1979); LCS uses Clopper–Pearson bounds (Clopper & Pearson, 1934) on pooled counts.

Model	Carlini	SA-MIA	SimMIA	LCS (Ours)
GPT-Neo 125M	0.089 [0.063,0.114]	0.168 [0.147,0.197]	0.115 [0.097,0.139]	**0.302** [0.276,0.328]
GPT-Neo 1.3B	0.097 [0.076,0.125]	**0.187** [0.158,0.212]	0.139 [0.113,0.169]	0.174 [0.153,0.196]
GPT-Neo 2.7B	0.092 [0.077,0.125]	0.172 [0.147,0.210]	0.139 [0.107,0.163]	**0.174** [0.153,0.196]
Pythia 410M	0.100 [0.084,0.142]	0.166 [0.138,0.191]	0.106 [0.081,0.138]	**0.289** [0.264,0.315]
Pythia 1.4B	0.117 [0.096,0.134]	**0.181** [0.152,0.203]	0.105 [0.083,0.126]	0.158 [0.138,0.179]
Pythia 12B	0.127 [0.106,0.153]	**0.193** [0.158,0.222]	0.111 [0.086,0.135]	0.148 [0.129,0.169]
RWKV-4 430M	0.106 [0.089,0.145]	0.186 [0.160,0.220]	0.098 [0.052,0.138]	**0.213** [0.190,0.236]
RWKV-4 3B	0.120 [0.109,0.151]	0.197 [0.171,0.218]	0.095 [0.012,0.129]	**0.214** [0.192,0.238]
RWKV-4 14B	0.143 [0.124,0.162]	**0.178** [0.159,0.199]	0.015 [0.003,0.034]	0.051 [0.040,0.065]
OLMo 1B	0.066 [0.048,0.087]	0.069 [0.046,0.092]	0.017 [0.007,0.030]	**0.098** [0.076,0.125]
OLMo 7B	**0.087** [0.069,0.114]	0.067 [0.048,0.091]	0.008 [0.002,0.017]	0.085 [0.064,0.110]
Mean	0.104	0.160	0.086	**0.173**

E Uncovering the "Tells": Extended Qualitative Examples

We visualize context-induced token shifts (δ_j) on Pythia-410M (Biderman et al., 2023), comparing a memorized pre-training corpus (`pile_pubmed_abstracts`) (Gao et al., 2020) with a held-out one (`bbc_news_2025`). Green ($\delta_j = 1$) and red ($\delta_j = -1$) highlights mark positions where the context ($K = 3$) corrects or disrupts the prediction, respectively.

On seen documents (Figure 11), the high density of red tokens illustrates that context interferes with the exact-match priors of memorized text (Xie et al., 2024; Zawalski et al., 2026).

Seen Documents: `pile_pubmed_abstracts` *Context hurts: 11 helped (green), 41 hurt (red)*

A monoclonal antibody, MN13, that recognizes specifically a novel substance between the postacrosomal sheath and the overlying plasma membrane in the mammalian sperm head. Monoclonal antibody MN13 raised against mouse spermatozoa specifically recognizes the postacrosomal region of the sperm head in several mammalian species. Colloidal gold-immunoelectron microscopy of demembranated mouse spermatozoa indicated that the antigen is associated with the outer layer of the periodic substructure apparently linking the postacrosomal sheath to the overlying plasma membrane. The antigen

. . . . . .

Genetic and clinical features of p450 oxidoreductase deficiency. P450 oxidoreductase (POR) deficiency is an autosomal recessive disorder of steroidogenesis with multiple clinical manifestations. POR is the electron donor for all microsomal P450 enzymes , including the three steroidogenic enzymes P450c17 (17alpha-hydroxylase/17,20-lyase), P450c21 (21-hydroxylase), and P450aro (aromatase). Since the first description of POR mutations in 2004, about 50 patients

. . . . . .

Morphometrics of a neotropical sandfly subspecies, Lutzomyia carrerai thula. In spite of obvious and frequently observed morphological variation in geographic populations of sand flies, taxonomists are reluctant to create subspecies. In doing so they avoid confusion caused by the accumulation of new taxa, but they also obscure natural processes, some of them related to probable species emergence. We investigated here how far the partitioning of size, shape and form could be a relevant technique for exploring early evolutionary processes. We used a typical case of

. . . . . .

Effectively addressing the mid- and long-term needs of young people affected by the tsunami in Aceh: an on-site assessment. Within two months of the Asian tsunami, a team of four individuals conducted an assessment on the post-disaster needs of young people in Aceh Province. In addition to assessing current needs, the team examined the extent to which young people (aged 14-24) were involved in the planning and implementation of ongoing rebuilding and relief efforts. Finally, the team assessed the degree to which young people could be

. . . . . .

Effect of haloperidol and risperidone on amyloid precursor protein levels in vivo. The neurotoxic beta-amyloid peptide of Alzheimer's disease is formed from the amyloid precursor protein (APP), which is a member of an evolutionarily highly conserved gene family with significant functional importance. Because behavioral and psychiatric symptoms treated with antipsychotics may influence the course of the disease, we have investigated traditional and atypical antipsychotic drugs, administered through the intraperitoneal route, for their effects on rat cortical APP. Western-immunoblotting was utilized for semi-quant

Figure 11: Positions where adding context changes the next-token prediction for **seen** documents. Red highlights ($\delta_j = -1$) heavily dominate, showing that context interferes with the model's exact-match verbatim memorization.

Conversely, Figure 12 demonstrates the model's behavior on unseen documents. Because these recent news articles post-date the pre-training cutoff, the model has no exact-match priors to disrupt. Instead, it relies heavily on in-context learning (Brown et al., 2020). The provided prefix serves as a beneficial semantic anchor, leading to predominantly green shifts as the context actively aids the next-token predictions rather than competing against memorized weights.

Unseen Documents: `bbc_news_2025` *Context helps: 51 helped (green), 10 hurt (red)*

William shocked by death of ex-n anny's stepson in New Orleans attack
Edward P ett ifer has been identified by the Metropolitan Police as the British national who was killed in New Orleans on 1 January
The Prince of Wales has said he and Catherine are "shocked and saddened" by the death of his former nanny's stepson, Edward Pettifer, who was killed in the New Orleans attack. The 31-year-old was among 14 people killed in a vehicle attack on New Year's Day. His stepmother

. . . . . .

Far age distances himself from Musk on Robinson
Elon Musk (L) and Nig el Farage at Donald Trump's Mar-A-Lago home in Florida last month
Re form UK leader Nigel Farage has distanced himself from Elon Musk's support for jailed far-right activist Stephen Yaxley-Lennon, also known as Tommy Robinson. The tech multi-billionaire added his voice to calls to release Yaxley-Lennon, who was jailed in October after admitting contempt of court by

. . . . . .

Stormzy banned from driving after mobile phone use
The 31-year-old rapper already had six points on his licence
Stormzy has been banned from driving for nine months after an off-duty police officer caught him using a mobile phone behind the wheel of a Rolls-Royce in London. The 31-year-old rapper, whose real name is Michael Ebenazer Owuo Junior, pleaded guilty via a letter to driving the Rolls-Royce Wraith while using a device on Addison Road, west

. . . . . .

French ski er Cyprien Sarrazin has left intensive care after successful surgery to drain a bleed on his brain following a fall in training.
Sarrazin, 30, was airlifted to hospital after losing control while training for the World Cup downhill in Bormio, Italy, on Friday. He had surgery the same day to reduce an intracranial haematoma.
"His state of health is stable. He was discharged from intensive care and admitted to a care unit," the French Ski Federation posted on X on Thursday.
"

. . . . . .

The film Wicked has been a huge hit with fans, becoming one of the highest-grossing movies of 2024.
Adapt ed from the stage musical, it tells the origin story of Elphaba, better known as the Wicked Witch of the West from The Wizard of Oz.
Elphaba's wig was braided in the capital by a woman who came to London as a Congolese refugee and now runs a hair salon in Seven Sisters.
This article was amended to make clear the wig was made

Figure 12: Positions where adding context changes the next-token prediction for **unseen** documents. Green highlights ($\delta_j = 1$) dominate, demonstrating that semantic context actively aids the model through in-context learning.