OLMo-Detect: A Multi-Stage, Confounder-Controlled Benchmark for Membership Inference on Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Membership inference on large language models (LLMs) aims to determine whether a given text sample was included in an LLM's training data, without access to its training corpus. Despite recent progress, existing benchmarks suffer from three limitations: limited coverage of training stages, insufficient distributional alignment between members and non-members, and lack of rigorous filtering of non-members against the training corpus. To address these limitations, we propose OLMo-Detect, a multi-stage, confounder-controlled benchmark built upon the fully open OLMo 2 pipeline. OLMo-Detect spans pre-training, mid-training, and post-training, explicitly aligns members and non-members on three key axes, and rigorously filters non-members via infini-gram. To assess robustness to distribution shifts, we further introduce OLMo-Detect (Shifted), a variant where members are misaligned with non-members. We evaluate 15 unsupervised and 3 supervised membership inference attacks (MIAs) across the OLMo 2 family, finding that: (i) overall performance is limited: the best unsupervised and supervised MIAs both reach an AUC of only 0.68, and supervised MIAs degrade under cross-domain evaluation; (ii) MIA performance peaks at mid-training and is lower at pre-training and post-training, a pattern driven by data type rather than a stage effect: curated math data is far more detectable than other types; (iii) overall scores improve from 1B to 13B but plateau at 32B; and (iv) no unsupervised MIA is robust to distribution shifts, with AUCs shifting by up to 0.42. Finally, we find that our findings on OLMo 2 generalize to OLMo 3 and non-OLMo models.

1 Introduction

Membership inference on large language models (LLMs) aims to determine whether a given text sample was included in an LLM's training data, without access to its training corpus. Despite recent progress (Shi et al., 2023; Ye et al., 2024; Zhang et al., 2024a; Dong et al., 2024; Duan et al., 2024; Fu et al., 2025; Tao et al., 2025; Chang et al., 2025; Zhang et al., 2025; Kim et al., 2026), there is still a lack of **multi-stage, confounder-controlled benchmarks**.

As illustrated in Table 1, existing benchmarks suffer from three limitations: (i) **Limited coverage of training stages.** Modern LLM development (Grattafiori et al., 2024; OLMo et al., 2025; Yang et al., 2025) has evolved into a three-stage pipeline of pre-training, mid-training, and post-training, where mid-training refines model capabilities through targeted high-quality data and annealing-style optimization (Tu et al., 2025). However, most existing benchmarks focus on the pre-training stage, with only a few targeting post-training. (ii) **Insufficient distributional alignment between members and non-members.** Most benchmarks either introduce distribution shifts, such as temporal drift (Shi et al., 2023; Ye et al., 2024; Zhang et al., 2024b; Fu et al., 2025) and fine-tuning-induced differences (Dong et al., 2024), or achieve only aggregate-level alignment by construction, such as train/held-out split (Duan et al., 2024; Tao et al., 2025), without explicitly controlling alignment on axes exploitable by existing membership inference attacks (MIAs). The only exception is OLMo-MIA (Kim et al., 2026), which aligns on semantic similarity. However, most MIAs (Carlini et al., 2021; Shi et al., 2023; Chang et al., 2023; Zhang et al., 2024a;b; Ye et al., 2024; Xie et al., 2024; Chang et al., 2025; Golchin & Surdeanu, 2025) exploit surface-form signals that manifest through exact wording rather than meaning, so semantically aligned data may still exhibit lexical shifts. (iii)

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.

Benchmark	Stage(s)	Domain(s)	Member	Non-Member	Assessments		
					C	A	F
WikiMIA (Shi et al., 2023)	Pre-training	Wikipedia	Pre-2017 data	Post-2023 data	✗	✗	✗
StackMIA (Ye et al., 2024)	Pre-training	Stack Exchange	Pre-2017 data	Post-2023 data	✗	✗	✗
DETCON (Dong et al., 2024)	Post-training	HumanEval, GSM8K	Outputs of post-SFT LLMs	Outputs of base LLMs	✗	✗	✗
MIMIR (Duan et al., 2024)	Pre-training	The Pile: 7 domains	Train splits	Held-out data	✗	✗	✓†
PatentMIA (Zhang et al., 2024b)	Pre-training	Google Patents	Pre-2023 data	Post-2024 data	✗	✗	✗
WikiMIA-24 (Fu et al., 2025)	Pre-training	Wikipedia	Pre-2017 data	Post-2024 data	✗	✗	✗
RL-MIA (Tao et al., 2025)	Post-training	AIME24/25, K&K, SAT	Training data	Held-out data	✗	✗	✗
OLMoMIA (Kim et al., 2026)	Pre-training	Dolma	Pre-checkpoint data	Post-checkpoint data	✗	✓‡	✗
OLMO-DETECT (ours)	Pre-training, Mid-training, Post-training	9 domains of OLMo 2	Distribution-matched training data	Distribution-matched, rigorously filtered non-members	✓	✓	✓

Table 1: Comparison of existing benchmarks for membership inference on LLMs. We assess on three criteria: (i) **Coverage (C)** of training stages (pre-training, mid-training, and post-training); (ii) **Explicit Alignment (A)** of distributions between members and non-members; (iii) **Rigorous Filtering (F)** of non-members against the training corpus. ✓† indicates that MIMIR uses a filter with a permissive threshold of 80%, leaving potentially high residual overlap. ✓‡ means that OLMoMIA aligns on semantic similarity, but most MIAs target exact wording rather than meaning.

Lack of rigorous filtering of non-members against the training corpus. Most benchmarks construct non-members without filtering out those sharing substantial n-gram overlap with the model's training data. MIMIR (Duan et al., 2024) is the sole exception but uses a 13-gram filter with a permissive threshold of 80%, leaving its non-members with potentially high residual overlap.

To address these limitations, we propose OLMO-DETECT, a multi-stage, confounder-controlled benchmark built upon the fully open training pipeline of OLMo 2 (OLMo et al., 2025). Specifically: (i) it spans pre-training, mid-training, and post-training, enabling systematic analysis of how detection performance varies across training stages; (ii) it explicitly aligns members and non-members on three key axes to minimize distribution shifts: text quality, temporal range, and lexical similarity; and (iii) it filters non-members against OLMo 2's training data via infini-gram (Liu et al., 2024)-based 13-gram filtering with a strict threshold of 20%, leaving non-members with an average residual overlap of only 5.31%. Moreover, we introduce OLMO-DETECT (SHIFTED), a benchmark variant where members are misaligned with non-members, to assess robustness to distribution shifts.

We evaluate 15 unsupervised and 3 supervised MIAs across the OLMo 2 family (1B/7B/13B/32B) on OLMO-DETECT, finding that: (i) **Overall performance is limited.** The best unsupervised method is Zlib (Carlini et al., 2021), one of the simplest MIAs, while the best supervised method is Sup-CAMIA (Chang et al., 2025); both achieve an average AUC of 0.68. Moreover, supervised MIAs degrade under cross-domain evaluation, indicating that learned signals do not transfer well. (ii) **Detectability is driven by data type rather than training stage.** Overall performance peaks at mid-training and is lower at pre-training and post-training, but this ordering reflects data type rather than a stage effect: curated math data (e.g., GSM8K (Cobbe et al., 2021)) is considerably more detectable than other types. (iii) Overall scores improve from 1B to 13B but plateau at 32B, suggesting that **scale alone may not continue to make detection easier.** Additionally, we evaluate unsupervised MIAs on OLMO-DETECT (SHIFTED) and find that **no unsupervised MIA is robust to distribution shifts**, with AUCs shifting by up to 0.42. Finally, we evaluate the top five unsupervised MIAs on OLMo 3 (7B/32B) (Olmo et al., 2026) and two non-OLMo families, DCLM-Baseline (1B/7B) (Li et al., 2025) and SmolLM2-1.7B (Allal et al., 2025), on the OLMO-DETECT domains they were trained on, and find that **our main findings generalize beyond OLMo 2**.

2 RELATED WORK

As summarized in Table 1, existing benchmarks for membership inference on LLMs can be grouped into four categories based on how members and non-members are constructed: (i) temporal-cutoff (Shi et al., 2023; Ye et al., 2024; Zhang et al., 2024b; Fu et al., 2025); (ii) partition-based (Duan et al., 2024; Tao et al., 2025); (iii) checkpoint-based (Kim et al., 2026); and (iv) model-output-based (Dong et al., 2024). Detailed descriptions are provided in Appendix A.

Existing MIAs can be grouped into unsupervised and supervised methods based on whether they use labeled members or non-members. Unsupervised MIAs can be classified into four categories: (i) likelihood-based (Yeom et al., 2018; Carlini et al., 2021; Shi et al., 2023; Zhang et al., 2024a;b; Ye et al., 2024; Xie et al., 2024; Fu et al., 2024; Chang et al., 2025; Kim et al., 2026); (ii) perturbation-based (Mattern et al., 2023); (iii) prompt-based (Golchin & Surdeanu, 2023; Chang et al., 2023; Deng et al., 2024; Golchin & Surdeanu, 2025); and (iv) output-distribution-based (Dong et al., 2024; Tao et al., 2025). Supervised MIAs include a supervised variant of CAMIA (Chang et al., 2025) and tuning-based methods (Fu et al., 2025; Zhang et al., 2025). Details are provided in Appendix B.

3 Task Definition

Let $\mathcal{M}$ denote an LLM trained on an unknown corpus $\mathcal{D}$. Given a candidate text x, the task of membership inference on LLMs is to infer whether $x \in \mathcal{D}$ via a scoring function $s(x; \mathcal{M})$, thresholded at λ: prediction$(x; \mathcal{M}) = \mathbb{1}[s(x; \mathcal{M}) \geq \lambda]$, where 1 and 0 denote member and non-member, respectively. Following prior work (Zhang et al., 2024a), we assume grey-box access to $\mathcal{M}$: output statistics (loss, logits, token probabilities) are available, internal parameters (weights, gradients) are not. Some methods further assume the ability to fine-tune $\mathcal{M}$, described in Appendix B.

4 Benchmark Construction

4.1 Motivation for Using OLMo 2

Constructing benchmarks requires LLMs with open training data to verify members and filter non-members against the corpus. Among existing open-source LLMs, GPT-NeoX (Black et al., 2022) and Pythia (Biderman et al., 2023) are limited to pre-training; OLMo 1 (Groeneveld et al., 2024) additionally includes post-training, but only for its 7B variant, and lacks mid-training; only OLMo 2 (OLMo et al., 2025) and OLMo 3 (Olmo et al., 2026) provide complete access across all three stages. We construct OLMo-Detect based on OLMo 2, as it was the available model when we started this work. Nonetheless, we show that patterns found on OLMo 2 generalize to OLMo 3 (Section 5.5).

4.2 Training Data of OLMo 2

We summarize the training data of OLMo 2 as follows: (i) **pre-training**: DCLM-Baseline (web), StarCoder (code), peS2o and arXiv (academic papers), OpenWebMath (web math), Algebraic Stack (math proofs and code), and Wikipedia/Wikibooks (encyclopedic); (ii) **mid-training**: three reused from pre-training (filtered DCLM-Baseline, peS2o, and Wikipedia/Wikibooks) and three new domains: FLAN (instruction data), Stack Exchange (Q&A), and Math Mix (curated math); and (iii) **post-training**: curated data for supervised fine-tuning (SFT), direct preference optimization (DPO), and reinforcement learning with verifiable rewards (RLVR). Details are provided in Appendix C.

4.3 Benchmark Overview

As summarized in Table 2, OLMo-Detect comprises nine domains spanning three training stages of OLMo 2's training data. Mid-training domains reused from pre-training are excluded; four domains are additionally excluded: pre-training's Wikipedia/Wikibooks (no source-matched non-members available), pre-training's arXiv and Algebraic Stack (no 13-gram-filtered non-members survived), and mid-training's FLAN (no instance-length-matched members available).

4.4 Construction of Non-Members

For each domain, we construct non-members by (i) identifying source-matched candidate pools, (ii) sampling candidates with strategies differing by domain type, (iii) retaining those that pass the infini-gram-based 13-gram filter, and (iv) allocating token budgets. Details are provided in Appendix E.

Source-Matched Candidate Pools. We first identify domains for which non-members can be drawn from the same underlying source as OLMo 2's training data, from either held-out data or the remainder of the original source corpus not used to train OLMo 2. (i) **Domains with Held-Out Data.** Pre-training: peS2o and OpenWebMath; mid-training: Math Mix's GSM8K; post-training: RLVR's

Stage	Domain	Subsets	Split	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		NM	150	217,117	FastText: 0.01809 ± 0.00007	2018-01-19 ± 930d	0.07
			M	150	217,016	FastText: 0.01811 ± 0.00000	2018-01-18 ± 940d	
	peS2o		NM	268	216,981	—	2022-12-14 ± 11d	0.10
			M	268	217,230	—	2022-12-15 ± 13d	
	OpenWebMath		NM	151	217,183	—	2019-10-13 ± 853d	0.11
			M	151	217,252	—	2019-10-17 ± 864d	
	StarCoder	Python, Java	NM	156	217,156	GitHub Stars: 35.65 ± 66.66 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	0.15
			M	156	217,222	GitHub Stars: 36.15 ± 64.67 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.36 ± 0.05	—	
Mid-training	Math Mix	GSM8K	NM	1,040	217,163	—	—	0.04
			M	1,040	217,132	—	—	
	Stack Exchange	Stack Overflow, Math Stack Exchange	NM	430	217,208	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 1.99	2015-03-31 ± 1378d	0.08
			M	430	217,202	Question Vote: 3.52 ± 1.74 Answer Vote: 5.54 ± 1.94	2015-03-30 ± 1367d	
Post-training	SFT†	WildChat, Aya	NM	225	217,057	—	—	0.11
			M	225	217,159	—	—	
	DPO†	UF-WildChat	NM	113	217,045	—	—	0.09
			M	113	217,287	—	—	
	RLVR	CoT-GSM8K, CoT-MATH	NM	290	217,110	—	—	0.01
			M	290	217,025	—	—	

Table 2: Overall per-domain statistics of OLMO-DETECT. Within each domain, members and non-members are aligned on up to three axes: Text **Quality**, **Temporal** Range, and **Lexical** Similarity. Text Quality uses OLMo 2's metric and is enforced for quality-thresholded domains; Temporal Range is enforced when timestamps are available; and Lexical Similarity, measured by Jensen–Shannon divergence (lower is better), is enforced for all domains. "—" denotes uncontrolled axes. **NM**: non-member; **M**: member. **Subsets** lists constituent subsets; blanks indicate single-subset domains whose subset shares the domain's name. †SFT and DPO are constructed per model size due to differing training data; values shown are for 1B. Each domain is split into dev/test sets at approximately 1:9 sample and token ratios; only totals are shown. Full statistics are in Appendix D.

CoT-GSM8K and CoT-MATH. (ii) **Domains with the Remainder of Original Source Corpora.** Pre-training: DCLM-Baseline and StarCoder; mid-training: Stack Exchange; post-training: SFT, DPO, and RLVR's IFEval. Among them, three are obtained by quality-thresholding the original corpora: DCLM-Baseline (fastText score ≥ 0.01811), StarCoder (GitHub stars ≥ 2; top-1 word frequency $\leq 30\%$ and top-2 word frequencies $\leq 50\%$), and Stack Exchange (question votes ≥ 3 and answer votes ≥ 5); the rest are obtained via random sampling from the original source corpora. For domains with multiple sampleable subsets, we select two from the top-5 subsets by training provenance to capture within-domain diversity.

Domain-Dependent Sampling Strategies. We then sample candidates based on whether a domain is quality-thresholded. (i) **Random Sampling for Domains without Quality Thresholding.** For domains whose candidate pools come from either held-out data or the remainder of original source corpora without quality thresholding, we randomly sample candidate non-members. (ii) **Boundary Sampling for Quality-Thresholded Domains.** For quality-thresholded domains, random sampling would yield data of systematically lower quality than that used to train OLMo 2. We thus apply **boundary sampling**, selecting candidates closest to OLMo 2's quality threshold. Since these candidates are drawn from the rejected remainder, they fall immediately below the threshold.

Infini-Gram-based 13-Gram Filtering. Finally, we filter out candidates with substantial training-data overlap via infini-gram. Infini-gram (Liu et al., 2024) is a suffix-array-based indexing system that supports efficient n-gram search over massive text corpora for arbitrarily large n. For each model size, we index its full training corpus, and apply 13-gram filtering with a strict threshold of 20%, a strategy established by Xu et al. (2025), rather than the permissive 80% used by MIMIR. Specifically, for each candidate, we extract all 13-grams G and compute the fraction that appear in the index, $\eta = \frac{|\{g \in G : \mathrm{count}(g) > 0\}|}{|G|}$. A candidate is retained as a non-member if $\eta < 20\%$.

Token Budget Allocation. From the resulting non-members, we sample toward a uniform budget of approximately 217K tokens per domain, determined by Math Mix's non-member token count, and allocate this budget proportionally to training provenance for multi-subset domains.

	Pre-training				Mid-training		Post-training			
	DCLM-Baseline	OpenWebMath	peS2o	StarCoder	Math Mix	Stack Exchange	SFT	DPO	RLVR	Avg
Overlap	4.07%	8.33%	3.11%	9.31%	6.65%	5.71%	6.32%	2.20%	2.06%	**5.31%**
BoW AUC	0.49	0.52	0.48	0.50	0.52	0.48	0.51	0.49	0.51	**0.50**

Table 3: Per-domain verification results, averaged across model sizes. **Overlap**: average 13-gram overlap of each domain's non-members against OLMo 2's training corpus. **BoW AUC**: blind detectability, measured by the AUC of a BoW classifier. Per-model-size results are in Tables 22 and 24.

4.5 Construction of Members

Given the constructed non-members per domain, we construct members that are distributionally aligned with them, via (i) identifying candidate pools, (ii) determining the alignment axes, and (iii) constrained sampling using simulated annealing. We provide detailed descriptions in Appendix F.

Candidate Pools. For domains without quality thresholding, candidate pools are the corresponding OLMo 2 training data. For quality-thresholded domains, to obtain candidates matched in quality to their non-member counterparts, we apply boundary sampling: since members are drawn from data that passed OLMo 2's quality threshold, this selects candidates immediately above the threshold.

Alignment Axes. We identify three key axes for aligning members with non-members: (i) **text quality**, already enforced via boundary sampling for quality-thresholded domains; (ii) **temporal range**, enforced when timestamp data are available, since temporal drift can inflate detection scores (Duan et al., 2024); and (iii) **lexical similarity**, measured by Jensen–Shannon divergence (JSD) between their token-frequency distributions and enforced across all domains, as most MIAs target exact wording rather than meaning (Section 1), making lexical mismatch a direct confounder.

Constrained Sampling via Simulated Annealing. Given the candidates, we select members that match their non-member counterparts subject to two constraints: (i) alignment on axes beyond text quality; (ii) matched instance and token counts, eliminating average instance length as a confounder. Let B and A denote the sets of non-members and candidate members for a domain. We search for a member subset $A^* \subseteq A$ that minimizes a distributional loss $\mathcal{L}$ via simulated annealing (Lecchini-Visintini et al., 2007), which probabilistically accepts suboptimal moves to escape local minima:

$$\begin{aligned} A^* = \arg\min_{A' \subseteq A} \quad & \mathrm{JSD}(p_{A'}, p_B) + \mathbb{1}[\text{temporal metadata available}] \cdot \mathcal{L}_{\text{temp}} \\ \text{s.t.} \quad & |A'| = |B|, \quad \left|\textstyle\sum_{x \in A'} \mathrm{tok}(x) - \sum_{x \in B} \mathrm{tok}(x)\right| \le \delta, \end{aligned} \tag{1}$$

where $p_{A'}$ and p_B denote token distributions of the candidate member subset and non-members, respectively; $\mathrm{tok}(x)$ denotes the token count of instance x; δ is the token margin, which is set to 250; JSD enforces lexical alignment and is bounded in $[0, 1]$; and $\mathcal{L}_{\text{temp}}$, which is normalized to the same scale as JSD, aligns members and non-members' mean and standard deviation of timestamps.

4.6 Benchmark Verification

Residual Overlap of Non-Members. Table 3 shows that non-members of OLMo-Detect exhibit an average 13-gram overlap of 5.31% against OLMo 2's full training corpus, with a maximum of just 9.31%, confirming the effectiveness of our infini-gram-based filtering scheme.

Verification of Lexical Alignment. Since lexical similarity has no ground truth and is measured using a summary statistic (JSD), residual lexical shifts may still remain. To rule this out, we follow Das et al. (2025) and probe each domain with a *blind* attack: a bag-of-words (BoW) classifier that predicts membership from text alone, without access to the LLM. If residual lexical confounders exist, the AUC of BoW will exceed chance. As shown in Table 3, every domain attains a blind AUC within 0.02 of 0.50, confirming that members and non-members are lexically well aligned.

4.7 OLMo-Detect (Shifted)

To examine the effect of distribution shifts on detection robustness, we introduce OLMo-Detect (Shifted). For each domain, non-members remain unchanged, but members are randomly sampled to satisfy only the constraint of matched instance and token counts, without enforcing alignment

Method	1B				7B				13B				32B				Model-Avg			
	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All
PPL^{L}	0.54	0.64	0.61	0.60	0.55	0.72	0.68	0.65	0.56	0.79	0.66	0.67	0.58	0.80	0.68	0.69	0.56	0.74	0.66	0.65
Zlib^{L}	0.57	**0.67**	**0.66**	**0.63**	0.59	**0.76**	0.66	0.67	0.59	**0.83**	**0.67**	**0.70**	0.61	**0.85**	**0.69**	**0.72**	0.59	**0.78**	**0.67**	**0.68**
Lower^{L}	0.57	0.61	0.59	0.59	0.57	0.69	0.62	0.63	0.58	0.75	0.65	0.66	0.58	0.73	0.67	0.66	0.58	0.69	0.63	0.63
MinK^{L}	0.55	0.65	0.62	0.61	0.56	0.73	0.67	0.65	0.57	0.80	0.66	0.68	0.59	0.81	0.69	0.70	0.57	0.75	0.66	0.66
MinK++^{L}	0.48	0.53	0.59	0.54	0.50	0.54	0.62	0.56	0.51	0.59	0.59	0.57	0.54	0.56	0.63	0.57	0.51	0.56	0.61	0.56
DCPDD^{L}	**0.59**	0.61	0.59	0.60	**0.60**	0.76	0.66	**0.67**	0.61	0.83	0.64	0.69	**0.63**	0.83	0.67	0.71	**0.61**	0.76	0.64	0.67
PAC^{L}	0.58	0.65	0.61	0.61	0.59	0.73	0.60	0.64	**0.62**	0.81	0.61	0.68	0.63	0.80	0.63	0.69	0.60	0.75	0.61	0.66
ReCaLL^{L}	0.50	0.43	0.50	0.48	0.54	0.29	0.49	0.44	0.55	0.34	0.48	0.45	0.52	0.37	0.54	0.48	0.53	0.36	0.50	0.46
EMMIA^{L}	0.56	0.53	0.61	0.57	0.58	0.54	0.61	0.57	0.57	0.56	0.60	0.58	0.61	0.59	0.47	0.56	0.58	0.56	0.57	0.57
CAMIA^{L}	0.55	0.67	0.59	0.60	0.53	0.73	**0.72**	0.66	0.54	0.80	0.66	0.67	0.54	0.82	0.66	0.67	0.54	0.76	0.66	0.65
Neigh.^{P}	0.55	0.59	0.53	0.56	0.54	0.64	0.55	0.58	0.55	0.73	0.53	0.60	0.55	0.76	0.57	0.63	0.55	0.68	0.55	0.59
Guided^{Pr}	0.47	0.52	0.50	0.50	0.50	0.47	0.50	0.49	0.47	0.48	0.49	0.48	0.48	0.55	0.46	0.50	0.48	0.51	0.49	0.49
DCQ^{Pr}	0.52	0.50	0.46	0.50	0.41	0.45	0.40	0.42	0.49	0.59	0.48	0.52	0.56	0.58	0.47	0.54	0.50	0.53	0.45	0.49
CDD^{O}	0.51	0.51	0.51	0.51	0.50	0.54	0.50	0.52	0.51	0.60	0.50	0.54	0.50	0.62	0.49	0.54	0.51	0.57	0.50	0.52
SelfCrit^{O†}	—	—	0.46	0.46	—	—	0.50	0.50	—	—	0.44	0.44	—	—	0.48	0.48	—	—	0.47	0.47
Method-Avg	0.54	0.58	0.56	0.56	0.54	0.61	0.59	0.58	0.55	0.68	0.58	0.60	0.57	0.69	0.59	0.61	0.55	0.64	0.58	0.58

Table 4: AUCs of unsupervised MIAs on OLMO-DETECT, grouped by family: ^{L} likelihood-based, ^{P} perturbation-based, ^{Pr} prompt-based, and ^{O} output-distribution-based. **Pre/Mid/Post** macro-average AUC over each stage's domains, and **All** macro-averages all stages. **Model-Avg** averages across model sizes, and **Method-Avg** averages across methods. The best method per column is in bold. †SelfCrit is evaluated on post-training's DPO and RLVR. Per-domain results are in Appendix I.

on text quality, temporal range, or lexical similarity. Mid-training's Math Mix and post-training's RLVR are excluded from OLMO-DETECT (SHIFTED), as randomly sampled members do not yield substantially different JSD from their corresponding non-members. Overall statistics of OLMO-DETECT (SHIFTED) are reported in Table 30, with detailed descriptions in Appendix G.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

We describe our experimental settings below, with details in Appendix H. (i) **Models.** We evaluate four instruction-tuned OLMo 2 models: 1B, 7B, 13B, and 32B. (ii) **Compared Methods.** We evaluate 15 unsupervised MIAs across four families: (a) likelihood-based: Perplexity (PPL) (Yeom et al., 2018), Zlib and Lowercase (Lower) (Carlini et al., 2021), Min-K% (MinK) (Shi et al., 2023), Min-K%++ (MinK++) (Zhang et al., 2024a), DCPDD (Zhang et al., 2024b), PAC (Ye et al., 2024), ReCaLL (Xie et al., 2024), EMMIA (Kim et al., 2026), and CAMIA (Chang et al., 2025); (b) perturbation-based: Neighborhood Attack (Neigh.) (Mattern et al., 2023); (c) prompt-based: Guided Instruction (Guided) (Golchin & Surdeanu, 2023) and DCQ (Golchin & Surdeanu, 2025); and (d) output-distribution-based: CDD (Dong et al., 2024) and Self-Critique (SelfCrit) (Tao et al., 2025) (evaluated only on DPO and RLVR, following its original setup). We also evaluate three supervised MIAs: the supervised variant of CAMIA (Sup-CAMIA) (Chang et al., 2025), MIA-Tuner (Fu et al., 2025), and FSD (Zhang et al., 2025). (iii) **Evaluation Metrics.** Following prior work (Shi et al., 2023), we report AUC and TPR@5%FPR. (iv) **Score Aggregation.** To aggregate subset-level scores into domain-level scores for multi-subset domains, domain-level scores into stage-level scores, and stage-level scores into final scores, we use macro-averaging, following Duan et al. (2024).

5.2 OVERALL PERFORMANCE

Unsupervised Methods. AUCs of unsupervised MIAs on OLMO-DETECT are reported in Table 4, with TPR@5%FPR results in Appendix J. Our main findings are: (i) **Overall performance is limited.** The best method, Zlib, reaches an overall AUC of only 0.68, while the average across methods is 0.58. This suggests that, under distribution-matched settings with minimal confounders, membership signals are substantially harder to detect than existing benchmarks suggest. Notably, Zlib, one of the earliest and simplest MIAs, outperforms every more sophisticated method. (ii) **Likelihood-based methods outperform other families.** Across families, likelihood-based MIAs generally perform best, with one exception: ReCaLL. ReCaLL assumes that a non-member prefix disrupts

Method	Setting	1B				7B				13B				32B				Model-Avg			
		Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All
Sup-CAMIA	In-domain	**0.56**	**0.72**	**0.61**	**0.63**	0.54	**0.78**	**0.72**	**0.68**	0.55	**0.86**	**0.72**	**0.71**	0.55	**0.84**	**0.67**	**0.69**	0.55	**0.80**	**0.68**	**0.68**
	LODO	0.56	0.63	0.52	0.57	**0.60**	0.73	0.62	0.65	**0.57**	0.66	0.55	0.59	**0.59**	0.70	0.62	0.64	**0.58**	0.68	0.58	0.61
MIA-Tuner	In-domain	0.54	0.60	0.53	0.56	0.53	0.61	0.55	0.56	0.54	0.60	0.53	0.56	0.54	0.53	0.51	0.52	0.54	0.59	0.53	0.55
	LODO	0.52	0.50	0.47	0.50	0.53	0.48	0.46	0.49	0.47	0.48	0.46	0.47	0.51	0.53	0.50	0.52	0.51	0.50	0.47	0.49

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.65	+0.00	−0.03	+**0.01**	−0.03	−0.01	−0.04	−0.02	−0.02	−0.05	−0.03
FSD (Zlib)	0.68	−0.04	−0.04	−0.02	−0.07	−0.04	−0.07	−0.03	−0.04	−0.07	−0.06
FSD (Lower)	0.63	−0.03	−0.02	−0.03	−0.05	−0.03	−0.04	−0.02	−0.02	−0.05	−0.03
FSD (MinK)	0.66	−0.05	−0.12	−0.07	−0.09	−0.08	−0.06	−0.03	−0.03	−0.06	−0.05

Table 5: Supervised MIAs on OLMO-DETECT under **In-domain** and **LODO** settings. **Top:** AUCs of Sup-CAMIA and MIA-Tuner; the best method per column is in bold. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members; we therefore report its AUC gain (ΔAUC) over the base MIA (in parentheses), averaged across stages. **Base AUC**: base MIA's AUC averaged across model sizes; positive gains are bolded. Details are in Appendix K.

predictions on members more than on non-members, but this effect is inverted on GSM8K, CoT-GSM8K, and CoT-MATH (AUC 0.10, 0.09, and 0.45; Tables 41 and 49): their rigid format turns the prefix into an in-context demonstration, improving member predictions instead of disrupting them. (iii) **Token-level normalization hurts performance.** MinK++, which applies token-level vocabulary normalization to MinK, performs markedly worse (0.56 vs. 0.66), suggesting that fine-grained normalization removes part of the membership signal under distribution-matched settings.

Supervised Methods. We evaluate supervised MIAs in two settings: (i) **In-domain**, trained on the evaluated domain's dev set, which measures the gain from same-distribution supervision over unsupervised MIAs; and (ii) **LODO** (leave-one-domain-out), trained on the other eight domains' dev sets and evaluated on the held-out domain, which measures cross-domain transfer. AUCs of supervised MIAs on OLMO-DETECT are reported in Table 5, with TPR@5%FPR results in Appendix L. Our main findings are: (i) **In-domain supervision adds little over unsupervised methods.** Under the **In-domain** setting, the best supervised method, Sup-CAMIA, reaches an overall AUC of 0.68, matching the best unsupervised MIA, Zlib (0.68), and exceeding its unsupervised counterpart CAMIA (0.65) by only 0.03; MIA-Tuner remains near chance (0.55), and FSD lowers the AUC of every base MIA on average. This indicates that matched-distribution labels add little beyond unsupervised membership signals. To rule out the small dev set as a confounder, we hold out 50% of each domain for evaluation and vary the supervision budget from 10% to 50% (Appendix M, Table 83). We find that increasing the budget improves Sup-CAMIA by only 0.04–0.06 AUC, confirming that the limited performance is not an artifact of the labeled-data budget. (ii) **Learned signals do not transfer well across domains.** Compared with **In-domain**, all supervised MIAs perform worse or no better under **LODO**: Sup-CAMIA drops to 0.61 (-0.07), MIA-Tuner to 0.49 (-0.06), and FSD's average gain over its base MIAs remains -0.04. This suggests that supervised MIAs largely learn domain-specific statistics rather than transferrable membership signals.

Domain-Unknown Setting. To simulate a real-world setting in which an auditor is unaware of a text's domain, we compute a single AUC over members and non-members pooled with equal instance counts across all domains, retraining supervised MIAs on the pooled dev sets. A full discussion is provided in Appendix N. In summary, unsupervised and supervised MIAs degrade slightly, by 0.02 and 0.03 AUC respectively, with the supervised drop concentrated in Sup-CAMIA (-0.06). The relative ranking is largely preserved (Spearman $\rho = 0.95$ over all 17 MIAs in Table 84).

5.3 PERFORMANCE ACROSS STAGES AND MODEL SIZES

Cross-Stage Performance. As shown in the **Model-Avg** columns of Tables 4 and 5, overall MIA performance peaks at mid-training and is considerably lower at post-training and pre-training. However, this pattern reflects data type rather than a stage effect: curated math data, i.e., instances drawn from purpose-built math datasets rather than harvested from the web, is far more detectable than any other data type. As the top half of Table 6 shows, the three curated math subsets, GSM8K, CoT-GSM8K, and CoT-MATH, achieve the highest AUCs across all MIAs, and their advantage is even

	Pre-training					Mid-training				Post-training				
	DCLM	peS2o	OWM	StarCoder		Math Mix	SE			SFT		DPO	RLVR	
				Python	Java	GSM8K	Math SE	SO		Aya	WildChat	UF-WildChat	CoT-GSM8K	CoT-MATH
AUC (all MIAs)	0.56	0.52	0.54	0.56	0.59	**0.71**	0.59	0.60		0.57	0.60	0.50	**0.70**	0.63
AUC (eff. MIAs)	0.62	0.50	0.57	0.59	0.64	**0.88**	0.61	0.61		0.59	0.62	0.50	**0.90**	**0.74**
Dup. count	9.00	1.12	1.00	1.00	1.00	6.00	6.00	6.00		2.09	2.00	1.13	1.25	1.50

	Stage (all subsets)			Stage (w/o curated math)			Curated math vs. others		
	Pre	Mid	Post	Pre	Mid	Post	Curated math	Others	Δ (#MIAs, p)
AUC (all MIAs)	0.55	0.65	0.59	0.55	0.59	0.54	0.69	0.56	+0.12 (13/16, 0.008)
AUC (eff. MIAs)	0.58	0.75	0.65	0.57	0.61	0.55	0.86	0.59	+0.27 (8/8, 0.004)

Table 6: Cross-stage analysis on OLMO-DETECT. **Top:** Per-subset AUCs averaged across model sizes and over all MIAs (excluding FSD, which reports gain, and SelfCrit, which is evaluated on DPO/RLVR) or *effective (eff.)* MIAs (overall AUC ≥ 0.60); bold means AUC > 0.70. *Dup. count*: training-corpus occurrences per member in each subset. DCLM: DCLM-Baseline; OWM: OpenWebMath; SE: Stack Exchange; SO: Stack Overflow. **Bottom:** Stage-level AUCs with and without the three curated math subsets (GSM8K, CoT-GSM8K, CoT-MATH), and AUCs on the curated math subsets versus the others. Δ: average per-MIA AUC gain on curated math subsets over the others; parentheses report the number of MIAs with positive gains and the one-sided Wilcoxon p-value.

more pronounced among effective MIAs (overall AUC ≥ 0.60), where all three exceed 0.70 and CoT-GSM8K reaches 0.90. Notably, the effect is specific to curated math data rather than to mathematical content in general: Math Stack Exchange (free-form forum math Q&A) and OpenWebMath (web math text) score only 0.61 and 0.57 among effective MIAs, in line with non-math subsets. To test this difference statistically, we conduct a one-sided Wilcoxon signed-rank test with $\alpha = 0.01$ on the paired per-MIA AUC differences between the curated math subsets and the others. As the bottom half of Table 6 shows, curated math subsets are significantly more detectable under both the all-MIA ($p = 0.008$) and the effective-MIA ($p = 0.004$) setting, and once they are excluded, the spread across stages shrinks substantially ($0.10 \rightarrow 0.05$ in the all-MIA and $0.17 \rightarrow 0.06$ in the effective-MIA setting), indicating that stage ordering is largely a composition effect. Finally, we examine duplication count, i.e., how many times each member is seen during training, as a potential confounder, and find that it does not explain the observed pattern: the most duplicated domain, DCLM, is less detectable than CoT-MATH, which has a duplication count of only 1.5.

Cross-Model-Size Performance. The **Method-Avg** row of Table 4 shows that the average AUC of unsupervised MIAs increases with model size, from 0.56 at 1B to 0.58 at 7B and 0.60 at 13B, but reaches only 0.61 at 32B. This trend is more pronounced for effective unsupervised MIAs (overall AUC ≥ 0.60): their average AUC rises from 0.61 at 1B to 0.68 at 13B, but gains only 0.01 more at 32B. Supervised methods (Table 5) even decline at 32B under the In-domain setting: Sup-CAMIA rises from 0.63 at 1B to 0.71 at 13B before falling to 0.69 at 32B, while MIA-Tuner remains at 0.56 through 13B and drops to 0.52 at 32B. The increase from 1B to 13B is consistent with MIMIR (Duan et al., 2024), which attributed a similar trend in Pythia models to greater overfitting. Beyond 13B, gains diminish and In-domain supervised performance declines, echoing Hayes et al. (2026), who report that MIA AUC does not increase monotonically with model size.

5.4 ROBUSTNESS TO DISTRIBUTION SHIFTS

To test the robustness of unsupervised MIAs to distribution shifts, we compare their per-domain AUCs on OLMO-DETECT (SHIFTED) and OLMO-DETECT, testing each difference using a paired instance-level bootstrap (Table 7). Our main findings are: (i) **No unsupervised MIA is robust.** Every MIA's AUC changes significantly on three to six of the seven domains, with each MIA's largest change ranging from 0.09 to 0.42 (mean 0.22); the largest changes concentrate in pre-training domains. (ii) **Whether distribution shifts inflate or degrade AUC follows no consistent pattern.** For all methods except ReCaLL and CDD, whose significant AUC changes are all inflations, there is no consistent shifting pattern across method–domain pairs: OpenWebMath inflates PPL's AUC by 0.29 but degrades Zlib's by 0.31; peS2o inflates CAMIA's AUC by 0.28 yet degrades EMMIA's by 0.17 (Table 85). Thus, shifted benchmarks distort AUC unpredictably. (iii) **Likelihood-based MIAs are most affected.** Eight of the ten likelihood-based MIAs change by 0.20–0.42 AUC, while the methods from other families (Neigh., Guided, DCQ, and CDD) change by at most 0.12.

	PPL	Zlib	Lower	MinK	MinK++	DCPDD	ReCaLL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided
I / D	4 / 2	2 / 2	5 / 1	2 / 1	1 / 3	2 / 3	4 / 0	4 / 1	5 / 1	5 / 1	3 / 2	2 / 1	3 / 0	1 / 2
Largest Δ	+0.29 (OWM)	−0.31 (OWM)	−0.30 (OWM)	+0.23 (OWM)	−0.16 (SC)	+0.16 (SC)	+0.24 (peS2o)	+0.28 (peS2o)	−0.20 (OWM)	+0.12 (SC)	+0.42 (OWM)	+0.09 (peS2o)	+0.12 (OWM)	+0.10 (SC)

Table 7: Effect of distribution shift on unsupervised MIAs (excluding SelfCrit), over the seven domains with shifted variants. **I / D**: number of domains on which the AUC increases / decreases significantly under the shift (Δ = OLMO-DETECT (SHIFTED) − OLMO-DETECT, averaged over model sizes; 95% paired instance bootstrap). **Largest** Δ: the largest change and the domain where it occurs. OWM: OpenWebMath; SC: StarCoder. Per-domain values are in Appendix O.

Domain	Comparison	PPL	Zlib	MinK	DCPDD	PAC	Avg
GSM8K	OLMo 2-7B / OLMo 3-7B	0.85 / 0.97	0.89 / 0.98	0.86 / 0.97	0.91 / 0.98	0.83 / 0.95	0.87 / 0.97
	OLMo 2-32B / OLMo 3-32B	0.98 / 0.99	0.99 / 1.00	0.97 / 0.99	0.98 / 0.98	0.92 / 0.98	0.97 / 0.99
DCLM-Baseline	OLMo 2-1B / DCB-1B	0.60 / 0.63	0.59 / 0.61	0.57 / 0.60	0.57 / 0.60	0.61 / 0.63	0.59 / 0.61
	OLMo 2-1B / SmolLM2-1.7B	0.60 / 0.62	0.59 / 0.62	0.57 / 0.59	0.57 / 0.61	0.61 / 0.64	0.59 / 0.62
OpenWebMath	OLMo 2-1B / DCB-1B	0.51 / 0.51	0.56 / 0.54	0.55 / 0.54	0.61 / 0.58	0.54 / 0.56	0.55 / 0.54
	OLMo 2-7B / DCB-7B	0.51 / 0.53	0.57 / 0.56	0.55 / 0.55	0.57 / 0.59	0.56 / 0.57	0.55 / 0.56

Table 8: AUCs of the five best-performing unsupervised MIAs on OLMo 2 and other model families.

5.5 GENERALIZATION BEYOND OLMO 2

To show that our findings are not artifacts of OLMo 2, we evaluate the five best-performing unsupervised MIAs (Zlib, DCPDD, MinK, PAC, and PPL) on OLMo 3 (Olmo et al., 2026) and two non-OLMo model families, comparing them with size-matched OLMo 2 models. (i) **OLMo 3.** The 7B and 32B instruction-following OLMo 3 models share the GSM8K subset of OLMO-DETECT's Math Mix domain with OLMo 2, which occurs at mid-training in both families. We further verify that OLMO-DETECT's GSM8K members are included in OLMo 3's training data while its non-members are excluded. As shown in Table 8, GSM8K remains highly detectable on OLMo 3 (AUC 0.95–1.00), with a method ranking similar to that on OLMo 2, PAC being the weakest in both families. While both families approach an AUC of 1.0 at 32B, OLMo 3 is more detectable at 7B, likely due to OLMo 3's stronger memorization capability. (ii) **Non-OLMo Models.** Eligible non-OLMo models must share training data with an OLMO-DETECT domain, whose non-members must either be excluded from the model's training data by construction based on clear documentation or be verifiable as absent, which requires open access to its training corpus. To the best of our knowledge, only two model families satisfy these strict conditions: DCLM-Baseline-1B/7B (DCB-1B/7B) (Li et al., 2025) and SmolLM2-1.7B (Allal et al., 2025). DCB-1B is trained on the full DCLM-Baseline corpus and OpenWebMath, covering two OLMO-DETECT domains, whereas DCB-7B is trained on an undisclosed subset of DCLM-Baseline, leaving only OpenWebMath valid for evaluation; SmolLM2-1.7B is trained on the full DCLM-Baseline corpus, covering the DCLM domain. Conveniently, the entire training data of DCB-1B/7B is a subset of OLMo 2's pre-training data, so the non-members of OLMO-DETECT's DCLM and OpenWebMath domains are clean for these models by construction; for SmolLM2, we apply the same 13-gram filter used in benchmark construction to each of its remaining training corpora, all of which are publicly released, and score both models on the non-members that are verified to be clean (Appendix P). As shown in Table 8, all three models closely track their size-matched OLMo 2 counterparts: every method is within 0.04 AUC, and both domains remain weakly detectable (AUC 0.51–0.64). Method rankings also transfer for all three models, with PAC performing best on DCLM and DCPDD best on OpenWebMath.

6 CONCLUSION

We introduced OLMO-DETECT, the first benchmark for membership inference on LLMs that spans pre-training, mid-training, and post-training, explicitly aligns members and non-members on three key axes, and rigorously filters non-members against the training corpus. After extensive evaluation, we find that existing MIAs remain limited, with the best MIA reaching an AUC of only 0.68. Moreover, curated math data is substantially more detectable than any other data type. Importantly, the main patterns generalize to OLMo 3 and non-OLMo models, showing that our findings are not artifacts of OLMo 2. We hope OLMO-DETECT serves as a rigorous testbed for future MIAs.

AI USE STATEMENT

In this work, we used generative AI tools for the following purposes:

1. **To aid or polish writing.** We used generative AI tools to refine wording, grammar, and clarity and to check the consistency of terminology and citation formatting. All AI-suggested edits were reviewed by the authors, who take full responsibility for the final text.
2. **For research ideation or execution.** We used generative AI tools to refine and double-check our implementations of the 18 compared MIA methods. All AI-refined components were verified and tested for correctness by the authors before use in our experiments. Additionally, we used generative AI tools to provide feedback on the benchmark construction methodology and experiments. All feedback was critically evaluated by the authors, who made all final methodological and experimental decisions.
3. **To draft sections of the paper.** We used generative AI tools to draft Appendix C (Detailed Descriptions of OLMo 2 Training Data), which catalogs extensive factual details about OLMo 2's publicly documented training pipeline. The authors checked all factual claims in the draft against OLMo 2's released documentation and wrote the final version themselves based on the draft.

We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

OLMO-DETECT is constructed entirely from data that OLMo 2 was trained on (members) or data drawn from the same source distributions (non-members). We do not scrape or introduce data from sources beyond those in OLMo 2's publicly documented training pipeline. We do not perform additional PII screening ourselves, but two domains inherit filtering from their sources (OLMo et al., 2025): peS2o applies regex-based PII detection and removes documents containing substantial PII (Soldaini et al., 2024); SFT includes WildChat, which anonymizes names, contact information, and other identifiers through a dedicated pipeline before release (Zhao et al., 2024).

REPRODUCIBILITY STATEMENT

Code and Data Release. We release our benchmark and full evaluation code at an anonymized repository: `https://anonymous.4open.science/r/OLMo-Detect-A56F`.

Benchmark Construction. Benchmark construction is described in Sections 4.3–4.7, with detailed constructions of non-members and members provided in Appendices E and F, respectively. Overall statistics for OLMO-DETECT are reported in Table 2, with detailed per-domain statistics provided in Appendix D. Statistics for OLMO-DETECT (SHIFTED) are provided in Appendix G.

Method Implementation and Evaluation. The overall experimental settings, including evaluated models, compared methods, and evaluation metrics, are described in Section 5. Detailed implementation information is provided in Appendix H, including (i) each method's original public repository, tuned hyperparameters, and search spaces; (ii) determinism and statistical reporting; (iii) continuous scoring for categorical methods; and (iv) reasons for excluding certain existing MIAs.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo

Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024. URL https://arxiv.org/abs/2405.09673.

Stella Biderman, Hailey Schoelkopf, Quentin 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, 2023. URL https://arxiv.org/abs/2304.01373.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In Angela Fan, Suzana Ilic, Thomas Wolf, and Matthias Gallé (eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bigscience-1.9. URL https://aclanthology.org/2022.bigscience-1.9/.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021. URL https://arxiv.org/abs/2012.07805.

Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7299–7321, 2025.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, memory: An archaeology of books known to chatgpt/gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, 2023.

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.

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.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

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. IEEE, 2025.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *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, 2024.

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 *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, 2024.

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? *arXiv preprint arXiv:2402.07841*, 2024.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010, 2024.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Mia-tuner: Adapting large language models as pre-training text detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27295–27303, 2025.

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. *arXiv preprint arXiv:2308.08493*, 2023.

Shahriar Golchin and Mihai Surdeanu. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Transactions of the Association for Computational Linguistics*, 13:809–830, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan

Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj

Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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/.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

Jamie Hayes, Ilia Shumailov, Christopher A. Choquette-Choo, Matthew Jagielski, George Kaissis, Milad Nasr, Sahra Ghalebikesabi, Meenatchi Sundaram Mutu Selva Annamalai, Niloofar Mireshghallah, Igor Shilov, Matthieu Meeus, Yves-Alexandre de Montjoye, Katherine Lee, Franziska Boenisch, Adam Dziedzic, and A. Feder Cooper. Exploring the limits of strong membership inference attacks on large language models, 2026. URL https://arxiv.org/abs/2505.18773.

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. *arXiv preprint arXiv:2103.03874*, 2021.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023. URL https://arxiv.org/abs/2311.10702.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, April 2017.

Gyuwan Kim, Yang Li, Evangelia Spiliopoulou, Jie Ma, and William Yang Wang. Detecting training data of large language models via expectation maximization. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1115–1129, 2026.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Andrea Lecchini-Visintini, John Lygeros, and Jan Maciejowski. Simulated annealing: Rigorous finite-time guarantees for optimization on continuous domains. *Advances in neural information processing systems*, 20, 2007.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023. URL https://arxiv.org/abs/2305.06161.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Huanyu Liu, Jia Li, Hao Zhu, Kechi Zhang, Yihong Dong, and Ge Li. Saturn: Sat-based reinforcement learning to unleash language model reasoning. *arXiv e-prints*, pp. arXiv–2505, 2025.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel

Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, 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, 2022. URL https://arxiv.org/abs/2203.02155.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. In *International Conference on Learning Representations*, volume 2024, pp. 20357–20379, 2024.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, https://github.com/allenai/pes2o.

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/.

Yongding Tao, Tian Wang, Yihong Dong, Huanyu Liu, Kechi Zhang, Xiaolong Hu, and Ge Li. Detecting data contamination from reinforcement learning post-training for large language models. *arXiv preprint arXiv:2510.09259*, 2025.

Chengying Tu, Xuemiao Zhang, Rongxiang Weng, Rumei Li, Chen Zhang, Yang Bai, Hongfei Yan, Jingang Wang, and Xunliang Cai. A survey on llm mid-training. *arXiv preprint arXiv:2510.23081*, 2025.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL https://arxiv.org/abs/2411.12372.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2742–2785, 2025.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8671–8689, 2024.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A Smith, and Hannaneh Hajishirzi. Infini-gram mini: Exact n-gram search at the internet scale with fm-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24955–24980, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. Data contamination calibration for black-box llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10845–10861, 2024.

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. IEEE, 2018.

Hengxiang Zhang, Songxin Zhang, Bingyi Jing, and Hongxin Wei. Fine-tuning can help detect pretraining data from large language models. In *International Conference on Learning Representations*, volume 2025, pp. 60902–60921, 2025.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024a.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5263–5274, 2024b.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

A Detailed Descriptions of Existing Benchmarks for Membership Inference on LLMs

This section provides detailed descriptions of each benchmark surveyed in Section 2.

Temporal-Cutoff Benchmarks. The central difficulty in benchmark construction is obtaining reliable membership labels without access to the model's training corpus. Temporal-cutoff benchmarks solve this by exploiting the model's training cutoff date: texts published before the cutoff are treated as members, while texts published after it are treated as non-members. WikiMIA (Shi et al., 2023) first instantiated this strategy on Wikipedia event pages, treating pre-2017 articles as members and post-2023 articles as non-members. Subsequent benchmarks extend it along two axes: (i)

broadening the coverage, where StackMIA (Ye et al., 2024) applies the same January 2017 cutoff to Stack Exchange posts, treating those after May 2023 as non-members, and PatentMIA (Zhang et al., 2024b) targets Chinese-language patents, treating those published before January 2023 as members and those with priority dates after March 2024 as non-members, to evaluate Chinese LLMs released in that window; and (ii) **refreshing the cutoff**, where WikiMIA-24 (Fu et al., 2025) retains WikiMIA's original January 2017 members but updates its non-members to articles published after March 2024, extending applicability to more recent models. However, text naturally drifts over time in vocabulary, topic, and style, so pre- and post-cutoff splits differ in more than membership alone, entangling genuine membership signals with distribution shifts. As a result, high performance on them likely reflects the shift rather than genuine membership detection (Duan et al., 2024).

Partition-Based Benchmarks. While temporal-cutoff benchmarks infer membership from publication dates, partition-based benchmarks obtain exact labels from a known training corpus, treating instances included in training as members and held-out or later-introduced instances as non-members. MIMIR (Duan et al., 2024) partitions the documented corpora of open models, namely The Pile (Gao et al., 2020) and Dolma (Soldaini et al., 2024), spanning a broad range of domains from general web text to academic and code corpora. To decontaminate non-members, it applies a Bloom-filter–based 13-gram deduplication filter against the training corpus, using a permissive overlap threshold of $\leq 80\%$: a candidate non-member is discarded only if *more than* 80% of its 13-grams already appear among training data, so a document sharing up to 80% of its 13-grams with the training data is still retained as a "non-member." Because the threshold is permissive, MIMIR reports substantial residual overlap between members and non-members across all measured domains, consistently exceeding 30%. In contrast, OLMO-DETECT applies 13-gram filtering with a far stricter $\leq 20\%$ threshold, reducing residual overlap to 5.31% on average (Table 3). Rather than partitioning an existing corpus, RL-MIA (Tao et al., 2025) manufactures its partition to simulate contamination during RL post-training, drawing from AIME24, AIME25, K&K (Xie et al., 2025), and SAT (Liu et al., 2025). Specifically, for AIME24 and AIME25, half of each benchmark's items are injected into the OpenR1-Math-46K (Guha et al., 2025) RL training corpus as members, with the remainder held out as non-members, while for K&K and SAT, original training instances serve as members and additional items synthesized via the same generation procedure serve as non-members. However, similar to MIMIR, RL-MIA provides no explicit mechanism to verify or control distributional alignment between its members and non-members.

Checkpoint-Based Benchmarks. OLMoMIA (Kim et al., 2026) constructs members and non-members from the training checkpoints of OLMo-7B (Groeneveld et al., 2024): members are examples seen before step 100k, and non-members are examples introduced between steps 400k and 500k, leaving the intervening steps unused. OLMoMIA is the only existing benchmark to explicitly control distributional alignment between members and non-members, via semantic similarity. Specifically, it embeds members and non-members using a text encoder, NV-Embed-v2 (Lee et al., 2025), clusters each split separately with k-means, and selects cluster pairs based on their inter-cluster distance, producing settings that range from maximally separated to near-indistinguishable. However, most MIA methods exploit surface-form signals that manifest through exact wording rather than meaning, while semantic similarity does not entail lexical similarity. For example, two texts describing the same event using synonyms can lie close in NV-Embed-v2's embedding space while sharing few overlapping n-grams. Consequently, two instances considered well-aligned under OLMoMIA's construction may still differ substantially in surface-form statistics that MIAs actually exploit, leading to a lexical shortcut that can inflate performance independent of membership signals.

Model-Output-Based Benchmarks. Both preceding strategies label texts, model-output-based benchmarks instead label model outputs, constructing members and non-members from the generations of models with and without exposure to the evaluation data. DETCON (Dong et al., 2024) simulates membership during the supervised fine-tuning (SFT) stage by fine-tuning a base LLM via low-rank adaptation (LoRA) (Hu et al., 2021) on HumanEval (Chen et al., 2021) (code generation) and GSM8K (Cobbe et al., 2021) (mathematical reasoning), varying the mixing ratio, learning rate, and occurrence count of these benchmarks in the fine-tuning data. Outputs from the fine-tuned model are treated as members, while outputs from the corresponding base model are treated as non-members. However, since members and non-members come from different model states rather than a single model, differences arising from the fine-tuning procedure, such as LoRA's own effect on

output style independent of the injected content, may provide a shortcut to separate the data. Moreover, LoRA fine-tuning exhibits dynamics that diverge from full-parameter training (Biderman et al., 2024), so detection signals calibrated on DETCON's LoRA-simulated membership may not transfer to membership arising from real-world LLM training.

B Detailed Descriptions of Existing MIAs

This section provides a taxonomy of existing MIAs. Notably, we restrict our scope to those applicable in the gray-box setting defined in Section 3. These methods can be grouped into unsupervised and supervised methods based on whether they require labeled members and non-members.

Unsupervised Methods. We classify unsupervised methods into four categories based on the signal they exploit: likelihood-based, prompt-based, perturbation-based, and output-distribution-based.

LIKELIHOOD-BASED. These methods infer membership from the model's token-level likelihoods, exploiting the tendency of LLMs to assign higher probability to previously seen text. The earliest instantiation, Perplexity (Yeom et al., 2018), directly thresholds the model's loss on the candidate text, based on the observation that the gap between member and non-member loss reflects overfitting. However, raw likelihood conflates two causes: a text may receive high probability because it was memorized, or merely because it is inherently predictable. Subsequent methods differ primarily in what they calibrate the likelihood against to remove this confound. The first family calibrates against properties of the candidate text itself: Zlib and Lowercase (Carlini et al., 2021) normalize the model's perplexity by the text's zlib-compressed entropy and the perplexity of its lowercased variant, respectively; Min-K% (Shi et al., 2023) restricts scoring to the $k\%$ least-probable tokens, hypothesizing that unseen text contains outlier tokens with abnormally low probability, and its extension Min-K%++ (Zhang et al., 2024a) further normalizes each token's log-probability by the mean and standard deviation over the vocabulary at that position, exploiting the insight that maximum-likelihood training pushes member tokens towards local maxima of the output distribution; DCPDD (Zhang et al., 2024b) calibrates token probabilities against a token-frequency prior estimated from a reference corpus, computing the cross-entropy between the two distributions to avoid misclassifying unseen text composed of frequent tokens. The second family calibrates against augmented variants of the candidate: PAC (Ye et al., 2024) extends Min-K% by constructing adjacent samples via word-level augmentation and computing a polarized distance that contrasts near and far probability regions to expose overfitted training samples; although it shares the augmentation idea with the perturbation-based Neighborhood Attack below, its final score remains a Min-K%-style likelihood statistic, so we group it here. The third family calibrates against other models: Smaller Ref (Carlini et al., 2021) scores candidates by the ratio of the target model's perplexity to that of a smaller model from the same family, exploiting the fact that smaller models memorize less of the shared training distribution, and SPV-MIA (Fu et al., 2024) removes the need for an externally trained reference by fine-tuning one on text elicited from the target model itself through self-prompting, scoring membership via a probabilistic-variation signal computed over perturbed variants of the candidate. Finally, the fourth family calibrates against the conditioning context rather than the text or model: ReCaLL (Xie et al., 2024) measures the relative drop in conditional log-likelihood when the candidate is prefixed with known non-member text, based on the finding that such conditioning reduces the likelihood of members substantially more than that of non-members; EMMIA (Kim et al., 2026) addresses ReCaLL's dependence on prefix selection by iteratively refining prefix scores through expectation-maximization; CAMIA (Chang et al., 2025) extracts richer signals from the per-token loss sequence, conditioning on prefix properties such as length, token diversity, and prefix perplexity, and computing loss-dynamics statistics including slope, fluctuation, and outlier counts, motivated by the observation that member losses decrease more smoothly as context accumulates.

PERTURBATION-BASED. Rather than calibrating the likelihood of the candidate alone, perturbation-based methods treat the candidate's synthetic neighbors as the reference: the neighbors match the candidate in content and fluency, so any gap between the candidate and its neighbors isolates memorization of surface form from inherent predictability. Neighborhood Attack (Mattern et al., 2023) generates neighbors via masked-word substitution with a mask-filling model and scores membership by the difference between the candidate's log-likelihood and the mean log-likelihood of its neighbors, flagging texts whose likelihood stands out from their own neighborhood.

PROMPT-BASED. Rather than reading likelihoods, prompt-based methods probe memorization through the model's responses to prompts, differing primarily in how much they withhold and what they ask the model to do with the rest. Guided Instruction (Golchin & Surdeanu, 2023) withholds the most: by prompting the model with the dataset name, partition, and the initial segment of an instance, it measures whether the completion reproduces the entire withheld remainder, scoring the completion against the reference continuation with ROUGE-L (Lin, 2004) and BLEURT (Sellam et al., 2020) and additionally employing an LLM judge to identify exact or near-exact matches. Comparing this guided completion with one generated from a general instruction that omits dataset identifiers isolates the contribution of dataset-specific memorization. Narrowing the target from a free-form continuation to a single slot, Chang et al. (2023) employ a name-cloze formulation in which a proper name is masked in a book passage and the model is asked to recover it without additional context, making successful recovery indicative of memorization of the source text, while TS-Guessing (Deng et al., 2024) adapts the same cloze principle, masking an incorrect answer option in a multiple-choice question and flagging membership when the exact missing option is reproduced. This narrowing reflects a trade-off: free-form completion probes more of the memorized content but requires soft matching, whereas slot filling admits exact-match verification at the cost of probing only a fragment. At the far end of this spectrum, DCQ (Golchin & Surdeanu, 2025) withholds nothing and shifts from production to recognition: it casts membership inference as a five-way multiple-choice quiz in which the original candidate appears alongside four word-level paraphrases that preserve its semantics. Since all options are semantically equivalent, a consistent preference for the original is attributable to memorization of surface form rather than content understanding.

OUTPUT-DISTRIBUTION-BASED. These methods share a common premise: membership leaves a fingerprint of abnormal concentration in the model's output distribution. CDD (Dong et al., 2024) reads this fingerprint at the string level, sampling multiple outputs for the same input and measuring the peakedness of their distribution around the greedy output via edit distance, flagging membership when the outputs are abnormally concentrated. Self-Critique (Tao et al., 2025) reads the same fingerprint in a setting where it is sharpest: membership introduced during RL post-training. Motivated by the observation that RL collapses the model's output entropy onto narrow reasoning paths, it prompts the model to produce two distinct responses to the same problem and flags membership when the two exhibit high similarity in the entropy space.

Supervised Methods. We classify supervised methods into two categories based on how they make use of labeled members and non-members: likelihood-based and tuning-based.

LIKELIHOOD-BASED. A supervised variant of CAMIA (Chang et al., 2025) uses labeled members and non-members to train a downstream classifier, leaving the target model untouched. It represents each candidate as a feature vector of its per-token loss-dynamics signals and trains a logistic regression classifier to combine them, applying PCA within each signal group to control dimensionality. The underlying likelihood-derived signals are unchanged, only their combination is learned.

TUNING-BASED. These methods use labeled data to tune the target model, differing in what is tuned and what the tuning is used to achieve. MIA-Tuner (Fu et al., 2025) keeps the model weights frozen and tunes a soft prompt injected into the target model, using both members and non-members, so that the tuned model becomes the detector directly: an instruction-tuning objective trains aligned models to answer *Yes* or *No* to a membership query, while a contrastive objective trains unaligned models, which cannot follow such instructions, to amplify the loss gap between members and non-members. FSD (Zhang et al., 2025) instead fine-tunes the model weights and is motivated by an empirical finding: briefly fine-tuning the target model on a small set of non-members disproportionately lowers non-members' scores under an existing method (e.g. Perplexity, Min-K%), while leaving members' scores largely unchanged. Rather than tuning the model to become the detector itself, as MIA-Tuner does, FSD uses this asymmetric shift as its detection signal: membership is inferred from a candidate's score deviation before versus after fine-tuning, and it requires only non-members for tuning.

Domain	Type	Tokens	Words	Bytes	Docs
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder (filtered)	Code	83.0B	70.0B	459B	78.7M
peS2o	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs/code	11.8B	10.8B	44.0B	2.83M
Wikipedia/Wikibooks	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

Table 9: Composition of OLMo 2 Mix 1124, the pre-training corpus of the OLMo 2 suite, reproduced from Table 4 of OLMo et al. (2025).

C Detailed Descriptions of OLMo 2 Training Data

This appendix provides a detailed breakdown of OLMo 2's training data across the three stages described in Section 4.2. Token counts, document counts, and filtering details are reproduced from the original OLMo 2 paper (OLMo et al., 2025).

C.1 Pre-training: OLMo 2 Mix 1124

OLMo 2 Mix 1124 comprises seven domains drawn from three source corpora. From DCLM (Li et al., 2025), OLMo 2 uses DCLM-Baseline (web pages), the top 10% quality subset of deduplicated DCLM-RefinedWeb (a RefinedWeb (Penedo et al., 2023)-style heuristically filtered version of the raw DCLM-Pool corpus), selected using a fastText classifier (Joulin et al., 2017). From Dolma 1.7 (Soldaini et al., 2024), it incorporates five subsets: peS2o (Soldaini & Lo, 2023) (academic papers), arXiv (STEM papers; originally sourced from RedPajama (Weber et al., 2024)), OpenWebMath (Paster et al., 2024) (math web pages), Algebraic Stack (Azerbayev et al., 2023) (math proofs and code), and Wikipedia/Wikibooks (encyclopedic content). The seventh domain, StarCoder (Li et al., 2023) (code), is derived from permissively licensed GitHub repositories. StarCoder is filtered to retain only repositories with at least 2 GitHub stars; within the retained set, documents whose most-frequent word constitutes over 30% of tokens or whose top-2 most-frequent words constitute over 50% are discarded.

The full per-domain composition of OLMo 2 Mix 1124 is summarized in Table 9. The 1B and 7B models are pretrained on 4 trillion tokens of this mix (1 epoch), the 13B on 5 trillion (1.2 epochs), and the 32B on 6 trillion (1.5 epochs).

C.2 Mid-training: Dolmino Mix 1124

Mid-training is an intermediate annealing stage between pre-training and post-training, in which the pretrained model is further trained on a smaller, higher-quality data mixture under a decaying learning rate to improve performance on targeted skills such as math and reasoning. OLMo 2's mid-training mixture, Dolmino Mix 1124, comprises six domains organized into a High-Quality Subset and a Dolmino Math Mix. The High-Quality Subset contains five domains: filtered DCLM-Baseline, peS2o, Wikipedia/Wikibooks, FLAN (Wei et al., 2021), and Stack Exchange. Dolmino Math Mix (abbreviated as Math Mix in the main paper) is a heterogeneous mixture aggregating synthetic math data (TuluMath, generated with GPT-4o (OpenAI et al., 2024); DolminoSynthMath and TinyGSM-MIND, generated with Qwen2.5-7B-Instruct (Qwen et al., 2025); and MathCoder2-Synthetic, collected from external sources and filtered using a FastText classifier trained on GPT-4o annotations), the GSM8K training split, OpenWebMath-filtered Metamath, and OpenWebMath-filtered CodeSearchNet. The full per-domain composition of Dolmino Mix 1124 is given in Table 10.

Within the High-Quality Subset, peS2o and Wikipedia/Wikibooks are reused from pre-training without modification. Filtered DCLM-Baseline applies additional filtering on top of DCLM-Baseline:

Domain	Type	Tokens	Words	Bytes	Docs
High-Quality Subset (five domains)					
Filtered DCLM-Baseline	High-quality web	752B	670B	4.56T	606M
FLAN	Instruction data	17.0B	14.4B	98.2B	57.3M
peS2o	Academic papers	58.6B	51.1B	413B	38.8M
Wikipedia/Wikibooks	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Stack Exchange	Q&A	1.26B	1.14B	7.72B	2.48M
High-quality total		**832.6B**	**739.8B**	**5.09T**	**710.8M**
Dolmino Math Mix (one domain, seven subsets)					
TuluMath	Synthetic math	230M	222M	1.03B	220K
DolminoSynthMath	Synthetic math	28.7M	35.1M	163M	725K
TinyGSM-MIND	Synthetic math	6.48B	5.68B	25.52B	17M
MathCoder2 Synth Books	Synthetic math	3.87B	3.71B	18.4B	2.83M
Metamath (OWM-filtered)	Math	84.2M	76.6M	741M	383K
CodeSearchNet (OWM-filtered)[†]	Code	1.78M	1.41M	29.8M	7.27K
GSM8K (training split)	Math	2.74M	3.00M	25.3M	17.6K
Dolmino Math Mix total		**10.7B**	**9.73B**	**45.9B**	**21.37M**

Table 10: Composition of Dolmino Mix 1124, the mid-training corpus of the OLMo 2 suite, reproduced from Table 5 of OLMo et al. (2025). The High-Quality Subset contributes five domains directly; the Dolmino Math Mix is itself a single domain whose composition is given by seven subsets. "OWM-filtered" indicates that the subset is further filtered using the OpenWebMath classifier designed by OLMo et al. (2025). [†]Although CodeSearchNet (OWM-filtered) is listed as part of the Dolmino Math Mix in OLMo et al. (2025), it was not used in the final training.

Domain	Tokens	50B sample		100B sample		300B sample	
		Domain %	Mix %	Domain %	Mix %	Domain %	Mix %
Filtered DCLM-Baseline	752B	3.23	47.2	6.85	50.2	20.78	51.9
FLAN	17.0B	50.0	16.6	100	16.7	200	11.3
Stack Exchange	1.26B	100	2.45	200	2.47	400	1.68
peS2o	58.6B	5.15	5.85	16.7	9.52	100	19.4
Wikipedia/Wikibooks	3.7B	100	7.11	100	3.57	400	4.86
Dolmino Math Mix	10.7B	100	20.8	200	17.5	400	10.8

Table 11: Per-domain composition of the three token-budget samples (50B, 100B, 300B) drawn from Dolmino Mix 1124 for mid-training, reproduced from Table 13 of OLMo et al. (2025). For each domain: "Domain %" is the fraction of that domain's full token count drawn into the sample, with values above 100 indicating repetition (e.g., 400 = 4× repeat); "Mix %" is the domain's share of the sample's total tokens (each Mix % column sums to 100%).

documents are first retained by the DCLM FastText classifier with threshold 0.0331 (retaining $\sim$ 65.6%), then restricted to FineWeb-Edu (Penedo et al., 2024) scores $\geq$ 2 ($\sim$ 20.3% retained overall). FLAN is decontaminated by removing training, validation, and test instances of evaluation benchmarks. Stack Exchange is taken from the 2024-09-30 dump, restructured into Q&A format, and filtered to retain only pairs with an accepted answer, at least 3 question votes, and at least 5 answer votes.

Three sample sizes (50B, 100B, 300B tokens) are drawn from the full mix, their per-domain compositions are summarized in Table 11. The 1B model anneals on a single 50B sample without averaging. The 7B, 13B, and 32B models apply *model soup* (Wortsman et al., 2022): multiple checkpoints are annealed on the same data mix using different randomized data orders, then averaged into the final midtrained model. The 7B averages three checkpoints trained on 50B samples, while the 13B and 32B each average four checkpoints (three trained on 100B samples and one on a 300B sample).

Subset	Count	# in SFT	Sampleable
General			
Tülu 3 Hardcoded	240	240	
OpenAssistant	88,838	7,132	✓
No Robots	9,500	9,500	
WildChat (GPT-4 subset)	241,307	100,000	✓
Knowledge Recall			
FLAN v2	89,982	89,982	
SciRIFF	35,357	10,000	✓
TableGPT	13,222	5,000	✓
Math and Reasoning			
Tülu 3 Persona MATH†	149,960	149,960	
Tülu 3 Persona GSM†	49,980	49,980	
Tülu 3 Persona Algebra	20,000	20,000	
OpenMathInstruct 2	21,972,791	50,000	✓
NuminaMath-TIR	64,312	64,312	
Coding			
Tülu 3 Persona Python	34,999	34,999	
Evol CodeAlpaca	107,276	107,276	
Safety and Non-Compliance			
Tülu 3 CoCoNot	10,983	10,983	
Tülu 3 WildJailbreak	50,000	50,000	
Tülu 3 WildGuardMix	50,000	50,000	
Multilingual			
Aya	202,285	100,000	✓
Precise Instruction Following			
Tülu 3 Persona IF	29,980	29,980	
Total	**23,221,012**	**939,344**	**6**

Table 12: Subsets of OLMo 2's SFT mix, reproduced from Table 7 of Lambert et al. (2025). "Count" is the total number of prompts available in the upstream source corpus; "# in SFT" corresponds to `tulu-3-sft-olmo-2-mixture` (7B and 13B; 939,344 prompts). The 1B and 32B models use `tulu-3-sft-olmo-2-mixture-0225` (866,138 prompts), which draws from the same 19 subsets but further removes date-cutoff instructions and applies majority voting over 5 completions to the †-marked subsets. The **Sampleable** column marks subsets with non-member candidates available for OLMo-Detect: ✓ = sampleable remainder available; blank = source corpus fully consumed (# in SFT = Count).

C.3 Post-training: Tülu 3 Recipe

The post-training stage follows the Tülu 3 recipe (Lambert et al., 2025) adapted to permissively-licensed sources, and proceeds in three sequential phases: SFT, DPO, and RLVR.

Supervised Fine-Tuning (SFT). Supervised fine-tuning (SFT) trains a pretrained language model on a collection of (prompt, response) pairs using next-token-prediction loss, teaching the

Stage	Domain	Subset	Prompts*	1B	7B	13B	32B	Sampleable
Post-training	DPO	SFT Reused	117,025	✓	✓	✓	✓	∼
		SFT IF	65,792	✓	✓	✓	✓	
		WildChat Unused‡	84,105	✓	✓	✓	✓	✓
		WildChat Reused‡	17,703	✓	✓	✓	✓	✓
		WildChat IF‡	10,794	✓		✓	✓	✓
		UltraFeedback (cleaned)	60,816	✓	✓	✓	✓	
		DaringAnteater IF	1,618	✓	✓	✓	✓	✓
		Tülu 3 Personas IF	19,890	✓	✓	✓	✓	✓
		Total prompts		378.3K	366.7K	377.7K	377.9K	**3**

Table 13: Prompt subsets for OLMo 2's DPO stage across all four model sizes. *Per-subset prompt counts are reproduced from Table 27 of OLMo et al. (2025) and apply to the 7B and 13B mixes; the 1B and 32B mixes draw from the same eight subset families with comparable but not identical per-subset counts. The 7B mix omits WildChat IF; chosen responses in the 1B and 32B mixes are additionally filtered to remove date-cutoff mentions. Subsets labeled "IF" contain prompts augmented with verifiable instruction-following constraints (IFEval-style). The **Sampleable** column marks subsets with uncontaminated candidates available for OLMO-DETECT: ✓ = sampleable remainder available; blank = source corpus fully consumed; ∼ = unknown (SFT Reused aggregates prompts from many SFT source datasets with varying sampleability, making it impossible to determine remainder without per-source attribution). ‡All three WildChat subsets draw from the same upstream source corpus and **together count as one of the three sampleable sources**; the remaining two are DaringAnteater IF and Tülu 3 Personas IF.

model to follow instructions and generate well-formatted responses. OLMo 2 draws its SFT prompts from the Tülu 3 SFT data sources (Lambert et al., 2025), organized into two mixes: `tulu-3-sft-olmo-2-mixture` (applied to 7B and 13B models), contains 939,344 prompts[1]; and `tulu-3-sft-olmo-2-mixture-0225` (applied to 1B and 32B models), contains 866,138 prompts and applies two additional prompt-level filters relative to the first mix: (i) prompts whose instruction text mentions an explicit date cutoff are removed, as training on such prompts caused the resulting model to hallucinate its own date cutoff and emit "As an AI language model..." prefaces at inference time; (ii) for the synthetic Persona MATH and Grade School Math subsets, five completions are sampled per prompt, and only prompts whose completions reach a majority answer are retained, reducing the rate of incorrect math answers absorbed during SFT. Both mixes inherit the Tülu 3 subset structure, as detailed in Table 12; the six subsets marked ✓: OpenAssistant, WildChat, SciRIFF, TableGPT, OpenMathInstruct 2, and Aya, are downsampled from larger upstream corpora.

Direct Preference Optimization (DPO). Direct preference optimization (DPO) (Rafailov et al., 2023) is a preference-alignment method that fine-tunes a language model directly on pairwise (chosen, rejected) data, bypassing the reward-model-based RL training loop used in conventional reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). DPO preference data is constructed for each model size via a common synthetic-data pipeline adapted from Tülu 3: (i) candidate responses are sampled from a pool of 20 external models together with on-policy data generated from intermediate OLMo 2 SFT checkpoints; (ii) GPT-4o-2024-08-06 serves as the judge, scoring responses along four dimensions (helpfulness, truthfulness, honesty, and instruction following); and (iii) scores are binarized using the Argilla procedure (Cui et al., 2024), which averages ratings across dimensions, selects the highest-scoring completion as the chosen response, and samples one of the remaining completions as the rejected response. All four mixes draw prompts from the eight Tülu 3 subset families listed in Table 13, yielding 378.3K prompts for the 1B, 366.7K for the 7B (excluding WildChat IF), 377.7K for the 13B, and 377.9K for the 32B. Chosen responses in the 1B and 32B mixes are additionally filtered to remove instances mentioning a date cutoff.

[1]The OLMo 2 paper (OLMo et al., 2025) reports 939,104 prompts for this mix, whereas the released HuggingFace dataset (`https://huggingface.co/datasets/allenai/tulu-3-sft-olmo-2-mixture`) contains 939,344 prompts. In this paper, we adopt the released dataset as the ground-truth count.

Stage	Component	1B	7B	13B	32B
Pre-training	OLMo 2 Mix 1124	✓	✓	✓	✓
Mid-training	Dolmino Mix 1124, 50B sample	×1	×3		
	Dolmino Mix 1124, 100B sample			×3	×3
	Dolmino Mix 1124, 300B sample			×1	×1
SFT	`tulu-3-sft-olmo-2-mixture-0225`	×2			×2
	`tulu-3-sft-olmo-2-mixture`		×2	×2	
DPO	Model-specific preference mix*	✓	✓	✓	✓
RLVR	Combined pass (GSM8K + MATH + IFEval)	✓	✓	✓	✓
	Additional GSM8K-only pass			✓	
	Additional MATH-only pass†	✓		✓	

Table 14: Per-model training corpora for infini-gram indexing. Entries denote inclusion in the model's indexed corpus, with multiplicities reflecting epoch and model-soup repetition (✓ = included once; $\times N$ = included N times). *Model-specific preference mixes: `olmo-2-0425-1b-preference-mix` (1B), `olmo-2-1124-7b-preference-mix` (7B), `olmo-2-1124-13b-preference-mix` (13B), `olmo-2-0325-32b-preference-mix` (32B). †The 1B model's additional MATH-only RLVR pass is inferred from the OLMo 2 training documentation in the `open-instruct` repository (`https://github.com/allenai/open-instruct/blob/main/docs/olmo2.md`).

Reinforcement Learning with Verifiable Rewards (RLVR). RLVR (Lambert et al., 2025) is a reinforcement learning fine-tuning method that replaces the learned reward model used in RLHF with a deterministic rule-based verifier. Completions receive a binary reward based on whether they can be programmatically verified as correct, e.g., through exact-match comparison against a ground-truth mathematical solution or satisfaction of explicit instruction-following constraints. The training data combines three subsets: the GSM8K training set, the MATH (Hendrycks et al., 2021) training set, and prompts from Lambert et al. (2025) containing verifiable instruction-following constraints (IFEval-style). GSM8K and MATH prompts are prepended with 8-shot and 3-shot chain-of-thought (CoT) in-context exemplars, respectively. The 7B and 13B models are trained with proximal policy optimization (PPO) (Schulman et al., 2017), with the value function initialized from a reward model trained on the same preference data used for DPO. The 1B and 32B models are trained with group relative policy optimization (GRPO) (Shao et al., 2024), which forgoes a reward model. All four models receive a single RLVR pass on the combined mix; the 13B model additionally undergoes two further passes, one on GSM8K alone and one on MATH alone, with the value function re-initialized from the reward model at each subsequent pass. The 1B model additionally undergoes one MATH-only pass (see Table 14 footnote).

D Detailed Per-Domain Statistics of OLMo-Detect

We provide detailed per-domain statistics for OLMo-Detect, as introduced in Section 4.3. Per-domain statistics for pre-training, mid-training, and post-training stages are reported in Tables 15–17, while per-subset statistics for domains comprising multiple subsets (StarCoder, Stack Exchange, SFT, and RLVR) are reported in Tables 18–21. For each domain, we partition the data into a development set (**dev**; for hyperparameter tuning) and a test set (**test**; for method evaluation), with both sample and token ratios set to approximately 1:9.

Interpreting the Lexical Column. JSD does not reach zero even between two samples drawn from the same source, because each sample misses some of the other's rare tokens. This floor rises as the token budget falls, and StarCoder Java has the smallest budget in the benchmark, 22.1K tokens against 184K–217K elsewhere. Splitting Java's non-members into two halves and comparing them against each other already yields 0.40, above the 0.30 reported for its members against its non-members. Restricting StarCoder Python to Java's token budget reproduces the effect, raising

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		NM	Dev	16	21,745	FastText: 0.01808 ± 0.00004	2018-01-18 ± 930d	0.07
				Test	134	195,372	FastText: 0.01809 ± 0.00008	2018-01-19 ± 930d	
				Total	**150**	**217,117**	FastText: 0.01809 ± 0.00007	2018-01-19 ± 930d	
			M	Dev	16	21,721	FastText: 0.01811 ± 0.00000	2018-01-17 ± 933d	
				Test	134	195,295	FastText: 0.01811 ± 0.00000	2018-01-18 ± 941d	
				Total	**150**	**217,016**	FastText: 0.01811 ± 0.00000	2018-01-18 ± 940d	
	peS2o		NM	Dev	27	21,793	—	2022-12-12 ± 10d	0.10
				Test	241	195,188	—	2022-12-14 ± 11d	
				Total	**268**	**216,981**	—	2022-12-14 ± 11d	
			M	Dev	27	21,804	—	2022-12-15 ± 14d	
				Test	241	195,426	—	2022-12-15 ± 13d	
				Total	**268**	**217,230**	—	2022-12-15 ± 13d	
	OpenWebMath		NM	Dev	15	21,761	—	2019-10-13 ± 851d	0.11
				Test	136	195,422	—	2019-10-13 ± 853d	
				Total	**151**	**217,183**	—	2019-10-13 ± 853d	
			M	Dev	15	21,793	—	2019-10-15 ± 851d	
				Test	136	195,459	—	2019-10-17 ± 865d	
				Total	**151**	**217,252**	—	2019-10-17 ± 864d	
	StarCoder	Python, Java	NM	Dev	16	21,536	GitHub Stars: 36.19 ± 67.09 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.37 ± 0.04	—	0.15
				Test	140	195,620	GitHub Stars: 35.59 ± 66.61 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	
				Total	**156**	**217,156**	GitHub Stars: 35.65 ± 66.66 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	
			M	Dev	16	21,587	GitHub Stars: 36.25 ± 65.18 Top-1 Freq: 0.28 ± 0.01 Top-2 Freq: 0.36 ± 0.04	—	
				Test	140	195,635	GitHub Stars: 36.14 ± 64.61 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.36 ± 0.05	—	
				Total	**156**	**217,222**	GitHub Stars: 36.15 ± 64.67 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.36 ± 0.05	—	

Table 15: Detailed per-domain statistics for the pre-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets; blank entries indicate single-subset domains whose subset shares the domain's name.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Math Mix	GSM8K	NM	Dev	103	21,755	—	—	0.04
				Test	937	195,408	—	—	
				Total	**1,040**	**217,163**			
			M	Dev	103	21,729	—	—	
				Test	937	195,403	—	—	
				Total	**1,040**	**217,132**			
	Stack Exchange	Stack Overflow, Math Stack Exchange	NM	Dev	43	21,760	Question Vote: 3.47 ± 2.20 Answer Vote: 5.12 ± 1.94	2015-03-27 ± 1378d	0.08
				Test	387	195,448	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 2.00	2015-03-31 ± 1378d	
				Total	**430**	**217,208**	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 1.99	2015-03-31 ± 1378d	
			M	Dev	43	21,845	Question Vote: 3.56 ± 2.18 Answer Vote: 5.37 ± 1.89	2015-03-29 ± 1302d	
				Test	387	195,357	Question Vote: 3.52 ± 1.68 Answer Vote: 5.56 ± 1.94	2015-03-29 ± 1374d	
				Total	**430**	**217,202**	Question Vote: 3.52 ± 1.74 Answer Vote: 5.54 ± 1.94	2015-03-29 ± 1367d	

Table 16: Sampling statistics for the mid-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets.

its value from 0.16 to 0.36. Java's 0.30 therefore reflects its size rather than residual misalignment, which the blind bag-of-words classifier confirms independently of sample size: it reaches an AUC of 0.51 on Java (Table 25), indistinguishable from chance.

Stage	Domain	Subset	Model	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	WildChat, Aya	1B, 32B	NM	Dev	21	21,697	—	—	0.11
					Test	204	195,360	—	—	
					Total	**225**	**217,057**			
				M	Dev	21	21,696	—	—	
					Test	204	195,463	—	—	
					Total	**225**	**217,159**			
			7B, 13B	NM	Dev	16	21,959	—	—	0.14
					Test	148	195,165	—	—	
					Total	**164**	**217,124**			
				M	Dev	16	21,738	—	—	
					Test	148	195,180	—	—	
					Total	**164**	**216,918**			
	DPO	UF-WildChat	1B	NM	Dev	11	21,705	—	—	0.09
					Test	102	195,340	—	—	
					Total	**113**	**217,045**			
				M	Dev	11	21,722	—	—	
					Test	102	195,565	—	—	
					Total	**113**	**217,287**			
			7B	NM	Dev	10	21,704	—	—	0.09
					Test	100	195,163	—	—	
					Total	**110**	**216,867**			
				M	Dev	10	21,699	—	—	
					Test	100	195,327	—	—	
					Total	**110**	**217,026**			
			13B	NM	Dev	10	21,730	—	—	0.12
					Test	100	195,321	—	—	
					Total	**110**	**217,051**			
				M	Dev	10	21,720	—	—	
					Test	100	195,539	—	—	
					Total	**110**	**217,259**			
			32B	NM	Dev	10	21,838	—	—	0.08
					Test	100	195,333	—	—	
					Total	**110**	**217,171**			
				M	Dev	10	21,779	—	—	
					Test	100	195,498	—	—	
					Total	**110**	**217,277**			
	RLVR	CoT-GSM8K, CoT-MATH	1B, 7B, 13B, 32B	NM	Dev	29	21,582	—	—	0.01
					Test	261	195,528	—	—	
					Total	**290**	**217,110**			
				M	Dev	29	21,543	—	—	
					Test	261	195,482	—	—	
					Total	**290**	**217,025**			

Table 17: Detailed per-stage statistics for the post-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets; blank entries indicate single-subset domains whose subset shares the domain's name.

E DETAILED DESCRIPTIONS OF NON-MEMBER CONSTRUCTION IN OLMO-DETECT

This appendix details the construction of non-members for each domain in OLMO-DETECT.

E.1 SOURCE-MATCHED CANDIDATE POOLS

Domains with Held-Out Data. These domains have a natural held-out split whose source already matches OLMo 2's training data.

Pre-training: (i) **peS2o**, which has two held-out files available, `validation-00000-of-00002.json.gz` and `validation-00001-of-00002.json.gz`. (ii) **OpenWebMath**, which has two held-out files available, `test.jsonl.zst` and `val.jsonl.zst`.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	StarCoder	Python	NM	Dev	10	19,364	GitHub Stars: 29.50 ± 61.41 Top-1 Freq: 0.31 ± 0.01 Top-2 Freq: 0.36 ± 0.04	—	0.16
				Test	90	175,594	GitHub Stars: 28.53 ± 59.41 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.06	—	
				Total	**100**	**194,958**	GitHub Stars: 28.63 ± 59.62 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.37 ± 0.05	—	
			M	Dev	10	19,551	GitHub Stars: 28.90 ± 58.22 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.35 ± 0.04	—	
				Test	90	175,551	GitHub Stars: 29.23 ± 57.87 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.36 ± 0.06	—	
				Total	**100**	**195,102**	GitHub Stars: 29.20 ± 57.90 Top-1 Freq: 0.29 ± 0.01 Top-2 Freq: 0.36 ± 0.06	—	
		Java	NM	Dev	6	2,172	GitHub Stars: 47.33 ± 74.29 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.38 ± 0.03	—	0.30
				Test	50	20,026	GitHub Stars: 48.28 ± 76.28 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.39 ± 0.03	—	
				Total	**56**	**22,198**	GitHub Stars: 48.18 ± 76.07 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.39 ± 0.03	—	
			M	Dev	6	2,036	GitHub Stars: 48.50 ± 73.75 Top-1 Freq: 0.28 ± 0.01 Top-2 Freq: 0.38 ± 0.03	—	
				Test	50	20,084	GitHub Stars: 48.56 ± 73.64 Top-1 Freq: 0.28 ± 0.02 Top-2 Freq: 0.36 ± 0.04	—	
				Total	**56**	**22,120**	GitHub Stars: 48.55 ± 73.65 Top-1 Freq: 0.28 ± 0.02 Top-2 Freq: 0.36 ± 0.04	—	

Table 18: Detailed per-subset statistics for StarCoder in the pre-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Stack Overflow	NM	Dev	37	18,489	Question Vote: 3.49 ± 2.26 Answer Vote: 5.14 ± 2.04	2015-02-06 ± 1381.39d	0.08
				Test	331	165,394	Question Vote: 3.48 ± 2.24 Answer Vote: 5.14 ± 2.05	2015-02-05 ± 1379.01d	
				Total	**368**	**183,883**	Question Vote: 3.48 ± 2.25 Answer Vote: 5.14 ± 2.05	2015-02-06 ± 1379.25d	
			M	Dev	37	18,489	Question Vote: 3.51 ± 2.23 Answer Vote: 5.32 ± 1.95	2015-02-06 ± 1329.25d	
				Test	331	165,471	Question Vote: 3.53 ± 1.66 Answer Vote: 5.58 ± 1.99	2015-02-06 ± 1378.98d	
				Total	**368**	**183,960**	Question Vote: 3.52 ± 1.72 Answer Vote: 5.55 ± 1.98	2015-02-06 ± 1374.07d	
		Math Stack Exchange	NM	Dev	6	3,271	Question Vote: 3.33 ± 1.80 Answer Vote: 5.00 ± 1.15	2016-01-22 ± 1323.12d	0.13
				Test	56	30,054	Question Vote: 3.25 ± 1.81 Answer Vote: 4.98 ± 1.64	2016-02-12 ± 1329.29d	
				Total	**62**	**33,325**	Question Vote: 3.26 ± 1.80 Answer Vote: 4.98 ± 1.60	2016-02-10 ± 1328.71d	
			M	Dev	6	3,356	Question Vote: 3.83 ± 1.86 Answer Vote: 5.67 ± 1.49	2016-02-03 ± 1074.89d	
				Test	56	29,886	Question Vote: 3.48 ± 1.81 Answer Vote: 5.45 ± 1.67	2016-01-28 ± 1302.79d	
				Total	**62**	**33,242**	Question Vote: 3.52 ± 1.82 Answer Vote: 5.47 ± 1.65	2016-01-28 ± 1282.52d	

Table 19: Detailed per-subset statistics for Stack Exchange in the mid-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Mid-training: **Math Mix**, whose GSM8K subset has one held-out file available, `test-00000-of-00001.parquet`. We additionally exclude the 200 instances designated by OLMo et al. (2025) as a development set for evaluation.

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Wildchat	NM	Dev	7	16,404	—	—	0.14
					Test	64	148,039	—	—	
					Total	**71**	**164,443**			
				M	Dev	7	16,438	—	—	
					Test	64	148,115	—	—	
					Total	**71**	**164,553**			
			Aya	NM	Dev	14	5,293	—	—	0.08
					Test	140	47,321	—	—	
					Total	**154**	**52,614**			
				M	Dev	14	5,258	—	—	
					Test	140	47,348	—	—	
					Total	**154**	**52,606**			
		7B, 13B	Wildchat	NM	Dev	7	16,715	—	—	0.18
					Test	66	148,305	—	—	
					Total	**73**	**165,020**			
				M	Dev	7	16,688	—	—	
					Test	66	148,362	—	—	
					Total	**73**	**165,050**			
			Aya	NM	Dev	9	5,244	—	—	0.06
					Test	82	46,860	—	—	
					Total	**91**	**52,104**			
				M	Dev	9	5,050	—	—	
					Test	82	46,818	—	—	
					Total	**91**	**51,868**			

Table 20: Detailed per-subset statistics for SFT in the post-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	RLVR	1B, 7B, 13B, 32B	CoT-GSM8K	NM	Dev	14	11,835	—	—	0.01
					Test	128	107,609	—	—	
					Total	**142**	**119,444**			
				M	Dev	14	11,796	—	—	
					Test	128	107,563	—	—	
					Total	**142**	**119,359**			
			CoT-MATH	NM	Dev	15	9,747	—	—	0.01
					Test	133	87,919	—	—	
					Total	**148**	**97,666**			
				M	Dev	15	9,747	—	—	
					Test	133	87,919	—	—	
					Total	**148**	**97,666**			

Table 21: Detailed per-subset statistics for RLVR in the post-training stage of OLMO-DETECT. **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Post-training: **RLVR**, whose held-out files for CoT-GSM8K and CoT-MATH subsets can be constructed by prepending the same CoT exemplars used in OLMo 2's RLVR setup (8-shot for GSM8K and 3-shot for MATH) to GSM8K's `test-00000-of-00001.parquet` and the test split of MATH's `train-00000-of-00001-7320a6f3aba8ebd2.parquet` (separated from the training portion during preprocessing), respectively.

Domains with the Remainder of Original Source Corpora. These domains lack a natural held-out split but are sampled from larger corpora. Therefore, candidate non-member pools are drawn from the portions of those original corpora that OLMo 2 did not use for training.

Pre-training: (i) **DCLM-Baseline.** DCLM-Baseline is the top 10% quality subset of deduplicated DCLM-RefinedWeb. From the remainder of DCLM-RefinedWeb, we randomly select three

shards as candidate pools (followed by deduplication): `global-shard_01-local-shard_0`, `global-shard_05-local-shard_6`, and `global-shard_07-local-shard_3`. (ii) **StarCoder**. OLMo 2's StarCoder is a quality-thresholded subset of the original StarCoder dataset (Li et al., 2023), retaining only data from repositories with at least 2 GitHub stars, where the most-frequent word constitutes at most 30% of tokens and the top-2 most-frequent words constitute at most 50%. The top-5 StarCoder subsets by training provenance are GitHub issues, Git commits, Markdown, Java, and Python, from which we select two distinct programming languages, Java and Python. From the rejected remainder of these two subsets, we consider the following files as candidate pools: Java: `train-00000-of-00087.parquet` through `train-00086-of-00087.parquet`; Python: `train-00000-of-00059.parquet` through `train-00058-of-00059.parquet`.

Mid-training: **Stack Exchange.** OLMo 2's Stack Exchange is a filtered version of 2024-09-30 dump, retaining only data with an accepted answer, at least 3 question votes, and at least 5 answer votes. The candidate non-member pool is therefore the rejected remainder of the original 2024-09-30 dump, processed using the same pipeline as OLMo 2 to obtain matched Q&A format.

Post-training: (i) **SFT.** Six subsets of OLMo 2's SFT are randomly downsampled from larger upstream corpora: OpenAssistant, WildChat, SciRIFF, TableGPT, OpenMathInstruct 2, and Aya (see Table 12 for details). Among them, we select two distinct subsets, WildChat and Aya. The candidate pools for SFT are therefore the unused portions of original WildChat (`train-00000-of-00014.parquet` to `train-00013-of-00014.parquet`) and original Aya (`test-00000-of-00001.parquet` and `train-00000-of-00001.parquet`). (ii) **DPO.** Five of OLMo 2's DPO prompt subsets are randomly sampled from larger corpora: three WildChat variants (Unused, Reused, and IF), DaringAnteater, and Tülu 3 Personas IF (see Table 13 for details). However, we only select WildChat (Unused) despite DPO's multi-subset nature. This is because neither DaringAnteater nor Tülu 3 Personas IF yields sufficient non-members after infini-gram-based 13-gram filtering; meanwhile, the three WildChat variants are not independently selectable either: since all three draw from the same upstream WildChat corpus, they share a single non-member remainder pool rather than three separate ones, and thus jointly count as one sampleable source. We therefore retain WildChat (Unused) alone, considering the remaining parts of original WildChat not used for OLMo 2 training (`train-00000-of-00014.parquet` to `train-00013-of-00014.parquet`) as candidate prompt pools. We then apply OLMo 2's UltraFeedback pipeline for response generation and preference annotation to obtain the source-matched candidate pools, yielding the final subset named UltraFeedback-WildChat (abbreviated as UF-Wildchat in the main paper). (iii) **RLVR.** The IFEval subset of OLMo 2's RLVR is randomly sampled from the Tülu 2 SFT mix (Ivison et al., 2023) and augmented with verifiable instruction-following constraints. Its candidate pool is therefore the remainder of the Tülu 2 SFT mix, augmented via the same constraint-construction pipeline.

RLVR Subset Selection. As illustrated above, RLVR has three sampleable subsets: CoT-GSM8K and CoT-MATH, drawn from held-out data, and IFEval, drawn from the remainder of the original source corpus. IFEval would add the most within-domain diversity, but we find that its non-members cannot be aligned lexically with members. Since alignment takes precedence over diversity, we select CoT-GSM8K and CoT-MATH, accepting that the two share a similar CoT construction.

Excluded Domains. Pretraining's Wikipedia/Wikibooks is excluded, as no source-matched candidate non-member pool is available.

E.2 Domain-Dependent Sampling Strategies

For domains whose candidate pools originate from either held-out files or the remainder of their source corpora without quality thresholding, sampling follows the random-sampling procedure described in Section 4.4. For quality-thresholded domains, we use boundary sampling to obtain candidate non-members whose quality is just below OLMo 2's selection threshold. Detailed descriptions are provided below.

DCLM-Baseline. We first identify the minimum fastText score in OLMo 2's DCLM-Baseline: 0.01811. Within the candidate pool, we then sample instances closest to this threshold, yielding over 350 candidate non-members (420K tokens) with an average fastText score of around 0.01809.

StarCoder. From the candidate pool, we sample instances whose top-1 word frequency is closest to 0.3 and top-2 word frequency is closest to 0.5, while enforcing the same GitHub-star constraint as OLMo 2, requiring each instance to have at least 2 stars. This yields 56 candidate non-members (22K tokens) for Java, with an average GitHub-star count of 74 and top-1/top-2 word frequencies of 0.30 and 0.39, and 250 candidate non-members (300K tokens) for Python, with an average GitHub-star count of 28 and top-1/top-2 word frequencies of 0.30 and 0.37. StarCoder is the only domain where the proportional allocation specified in Section 4.4 cannot be met: although Java constitutes a larger share of OLMo 2's training data than Python and would therefore be expected to contribute more tokens, too few Java candidates satisfy our filtering criteria to meet that target. We therefore retain all available Java non-members (56 instances, 22K tokens) and use Python to fill the remainder of the domain's $\sim$217K-token budget.

Stack Exchange. From the candidate pool, we sample instances satisfying either (a) question vote $= 2$ and answer vote ≥ 5, or (b) question vote ≥ 3 and answer vote $= 4$. This yields over 400 candidate non-members (200K tokens) for Stack Overflow, with an average question vote of 3.48 and answer vote of 5.14, and around 100 candidate non-members (50K tokens) for Math Stack Exchange, with an average question vote of 3.26 and answer vote of 4.98.

E.3 Infini-gram-Based 13-gram Filtering

After obtaining candidate non-members for each domain, we perform 13-gram filtering via infini-gram to retain those with minimal training-data overlap, as described in Section 4.4. Here, we detail the index construction: we build an infini-gram index over each OLMo 2 model's full training corpus, defined as the union of all training-stage documents with multiplicities reflecting epoch and model-soup repetition. The composition for each model size is summarized in Table 14.

Excluded Domains. Although source-matched candidate pools are identified for arXiv, using held-out files `arXiv_000.jsonl.zst` through `arXiv_099.jsonl.zst` (test) and `arXiv_000.jsonl.zst` through `arXiv_099.jsonl.zst` (validation), and for Algebraic Stack, using held-out files in its test and validation directories, no candidates from either pool survive 13-gram filtering. Therefore, both arXiv and Algebraic Stack are excluded.

E.4 Token Budget Allocation

We sample from non-members that survive 13-gram filtering, following the rules in Section 4.4. To ensure meaningful statistical analysis, each domain contains at least 100 non-members, with at least 50 instances per subset in multi-subset domains, to ensure meaningful statistical analysis.

F Detailed Descriptions of Member Construction in OLMo-Detect

This appendix details the construction of members for each domain in OLMo-Detect.

F.1 Candidate Pools

F.1.1 Domains without Quality Thresholding

For the following domains, candidate pools are the corresponding training data used by OLMo 2.

Pre-training: (i) **peS2o,** whose candidate member pool is OLMo 2's peS2o training data, `pes2o-0000.json.gz` through `pes2o-0025.json.gz`. (ii) **OpenWebMath,** whose candidate member pool is OLMo 2's OpenWebMath training data, `open-web-math-train-0000.json.gz` through `open-web-math-train-0012.json.gz`.

Mid-training: **Math Mix,** whose only subset, GSM8K, has a candidate member pool drawn from its corresponding OLMo 2 training data, `0.jsonl.zst`.

Post-training: (i) **SFT.** Both of SFT's subsets, WildChat and Aya, draw their candidate member pool from the corresponding OLMo 2 training data, filtered by source keyword: `tulu_v3.9_wildchat_100k` for WildChat and `tulu_v3.9_aya_100k` for Aya. The 1B and 32B models draw shards 0–5 of `tulu-3-sft-olmo-2-mixture-0225`; the 7B and 13B models draw shards 0–5 of `tulu-3-sft-olmo-2-mixture`. (ii) **DPO.** For each model size, candidate pools are drawn from OLMo 2's corresponding DPO training data for DPO's only subset, UltraFeedback-WildChat, filtered using a model-specific keyword: 1B draws shards 0–4 of `olmo-2-0425-1b-preference-mix` (keyword: `olmo2-1b-wildchat-unused-on-policy`); 7B draws shards 0–5 of `olmo-2-1124-7b-preference-mix` (keyword: `wildchat_v3.9_unused_on_policy_olmo2_7b`); 13B draws shards 0–5 of `olmo-2-1124-13b-preference-mix` (keyword: `wildchat_v3.9_unused_on_policy_olmo2_13b`); and 32B draws shards 0–4 of `olmo-2-0325-32b-preference-mix` (keyword: `wildchat_v3.9_unused_off_policy_prompts-olmo32`). (iii) **RLVR.** The candidate pools for both the CoT-GSM8K and CoT-MATH subsets of RLVR originate from the single shard of `RLVR-GSM-MATH-IF-Mixed-Constraints`, filtered using the keywords `gsm8k` and `MATH`, respectively.

F.1.2 Quality-Thresholded Domains

OLMo 2's training data in the following domains has systematically higher quality than the corresponding non-members, since only instances that passed OLMo 2's quality threshold were selected for training. To align text quality with non-members, we select training data closest to the threshold as the candidate pool.

DCLM-Baseline. We restrict the candidate pool to the lowest-scoring instances, those at the 10% quality threshold with a fastText score of 0.01811, drawn from three shards of OLMo 2's DCLM-Baseline: `global-shard_01-local-shard_0`, `global-shard_05-local-shard_6`, and `global-shard_07-local-shard_3`. This yields a candidate member pool of over 300 instances (580K tokens).

StarCoder. We restrict the candidate pool to instances whose top-1 frequency lies within 0.28–0.3 or top-2 frequency lies within 0.48–0.5. Regarding GitHub stars, since candidates are drawn from OLMo 2's StarCoder training data, the ≥ 2 GitHub-star constraint, enforced for non-members, is already satisfied by construction. This yields a candidate member pool of over 8,000 instances (7.3M tokens) for Java and over 5,500 instances (15.9M tokens) for Python.

Stack Exchange. We restrict the candidate pool to instances with question vote $= 3$ or answer vote $= 5$. This yields a candidate member pool of over 300K instances (200M tokens) for Stack Overflow and over 30K instances (19M tokens) for Math Stack Exchange.

F.2 Alignment Axes

We align members and non-members on three axes: text quality, temporal range, and lexical similarity, as discussed in Section 4.5. Each axis is either a signal that existing MIAs are known to exploit or a by-product of OLMo 2's filtering that must therefore be controlled. These three are not exhaustive. Adding more axes is possible but carries a cost: each further constraint makes the simulated-annealing search substantially harder and shrinks the feasible candidate pool, reducing subset sizes and therefore the reliability of the resulting statistical analysis. We thus stop at the three most important axes, prioritizing subset size and statistical reliability over additional constraints.

F.3 Constrained Sampling via Simulated Annealing

This section provides a detailed description of the simulated-annealing procedure used to sample members, introduced in Section 4.5. The procedure is applied independently to each domain, or to each subset within multi-subset domains, against its own corresponding non-member set.

	Pre-training				Mid-training		Post-training			
	DCLM-Baseline	OpenWebMath	peS2o	StarCoder	Math Mix	Stack Exchange	SFT	DPO	RLVR	**Avg**
1B	$4.07_{\pm 4.87}$	$8.33_{\pm 5.69}$	$3.11_{\pm 4.01}$	$9.31_{\pm 5.24}$	$6.65_{\pm 4.77}$	$5.71_{\pm 5.34}$	$6.08_{\pm 5.24}$	$2.12_{\pm 3.01}$	$2.06_{\pm 4.55}$	5.27
7B	$4.07_{\pm 4.87}$	$8.33_{\pm 5.69}$	$3.11_{\pm 4.01}$	$9.31_{\pm 5.24}$	$6.66_{\pm 4.75}$	$5.71_{\pm 5.34}$	$6.54_{\pm 5.29}$	$1.98_{\pm 3.05}$	$2.06_{\pm 4.55}$	5.31
13B	$4.07_{\pm 4.87}$	$8.33_{\pm 5.69}$	$3.11_{\pm 4.01}$	$9.31_{\pm 5.24}$	$6.65_{\pm 4.75}$	$5.71_{\pm 5.34}$	$6.55_{\pm 5.30}$	$2.52_{\pm 3.37}$	$2.06_{\pm 4.55}$	5.37
32B	$4.07_{\pm 4.87}$	$8.33_{\pm 5.69}$	$3.11_{\pm 4.01}$	$9.31_{\pm 5.24}$	$6.64_{\pm 4.75}$	$5.71_{\pm 5.35}$	$6.12_{\pm 5.31}$	$2.20_{\pm 3.15}$	$2.06_{\pm 4.55}$	5.28
Avg	4.07	8.33	3.11	9.31	6.65	5.71	6.32	2.20	2.06	**5.31**

Table 22: Per-model-size 13-gram overlap η (%; mean$_{\pm \text{std}}$ over non-members) for each domain against the training corpus of the corresponding OLMo 2 model size. Multi-subset domains are macro-averaged across subsets (Table 23), with the domain-level standard deviation computed as the root mean square of the subset-level standard deviations.

Setup. Given a candidate member pool A and a fixed non-member split with token distribution p_B, the goal is to select a subset $A' \subseteq A$ satisfying three constraints: (i) $|A'|$ equals the non-member instance count $|B|$; (ii) the total token count of A' is within $\delta = 250$ of the non-member token count; and (iii) the distributional loss $\mathcal{L}$ against the non-member reference is minimized:

$$\mathcal{L} = \mathrm{JSD}(p_{A'}, p_B) + \mathbb{1}[\text{temporal metadata available}] \cdot \mathcal{L}_{\text{temp}}, \tag{2}$$

where $p_{A'}$ is the token distribution of A', and JSD denotes the Jensen–Shannon divergence, computed with base-2 logarithms so that $\mathrm{JSD}(p_{A'}, p_B) \in [0, 1]$, enforcing lexical alignment. The temporal alignment term takes the form:

$$\mathcal{L}_{\text{temp}} = \frac{|\mu_{A'} - \mu_B| + |\sigma_{A'} - \sigma_B|}{\sigma_B}, \tag{3}$$

with μ and σ denoting the mean and standard deviation of the temporal axis (in days), normalized by the non-member standard deviation σ_B.

Algorithm. We initialize by randomly partitioning A into $(A', \text{leftover})$ such that $|A'|$ matches the target instance count $|B|$. The algorithm then runs for $T_{\max} = 10^7$ steps, with each step proposing a swap between one instance in A' and one in the leftover pool. The procedure operates in two phases. In the **greedy token-matching phase**, used when the current token count is outside the δ margin, the swap is accepted only if it brings the token count closer to target. In the **alignment optimization phase**, used once within margin, the swap is accepted if the post-swap token count remains within margin and the swap satisfies the Metropolis criterion: accepted if $\Delta\mathcal{L} < 0$, otherwise accepted with probability $\exp(-\Delta\mathcal{L}/T)$.

Temperature Schedule. The temperature T is initialized at $T_0 = 10^{-3}$ and decays geometrically: $T \leftarrow 0.999995 \cdot T$ after each accepted swap in the alignment optimization phase. This schedule provides broad exploration early in the search and progressive convergence as the temperature decreases.

Reproducibility. All simulated-annealing runs use a fixed random seed of 42, and the resulting subset A^* is retained as the final member split.

Excluded Domains. Mid-training's FLAN has sufficient candidate members and non-members, but no member and non-member pair can be matched on both instance count and total tokens, even with the token margin relaxed to 1,000. We therefore exclude it from OLMO-DETECT.

G DETAILED PER-DOMAIN STATISTICS OF OLMO-DETECT (SHIFTED)

This appendix provides detailed statistics of OLMO-DETECT (SHIFTED). Overall statistics are reported in Table 30; detailed per-domain statistics are reported in Tables 31–33; per-subset statistics for multi-subset domains, StarCoder, Stack Exchange, and SFT, are reported in Tables 34–36. Math Mix and RLVR are excluded, since random sampling does not yield members with meaningful distribution shifts in these domains.

Domain	Subset	1B	7B	13B	32B	Avg
StarCoder	Python	$7.17_{\pm 5.28}$	$7.17_{\pm 5.28}$	$7.17_{\pm 5.28}$	$7.17_{\pm 5.28}$	7.17
	Java	$11.45_{\pm 5.20}$	$11.45_{\pm 5.20}$	$11.45_{\pm 5.20}$	$11.45_{\pm 5.20}$	11.45
Stack Exchange	Math Stack Exchange	$5.67_{\pm 4.91}$	$5.67_{\pm 4.91}$	$5.67_{\pm 4.91}$	$5.68_{\pm 4.92}$	5.67
	Stack Overflow	$5.74_{\pm 5.74}$	$5.74_{\pm 5.74}$	$5.74_{\pm 5.75}$	$5.75_{\pm 5.74}$	5.74
SFT	WildChat	$5.16_{\pm 4.49}$	$4.53_{\pm 4.53}$	$4.55_{\pm 4.53}$	$5.14_{\pm 4.59}$	4.84
	Aya	$7.01_{\pm 5.89}$	$8.55_{\pm 5.95}$	$8.55_{\pm 5.96}$	$7.10_{\pm 5.95}$	7.80
DPO	Prompt	$0.00_{\pm 0.00}$	$0.13_{\pm 1.04}$	$0.00_{\pm 0.00}$	$0.01_{\pm 0.11}$	0.04
	Chosen	$3.55_{\pm 3.58}$	$3.06_{\pm 3.46}$	$4.20_{\pm 4.47}$	$3.73_{\pm 4.07}$	3.63
	Rejected	$2.80_{\pm 3.79}$	$2.75_{\pm 3.84}$	$3.37_{\pm 3.75}$	$2.85_{\pm 3.63}$	2.94
RLVR	CoT-GSM8K	$1.45_{\pm 3.77}$	$1.45_{\pm 3.77}$	$1.45_{\pm 3.77}$	$1.45_{\pm 3.77}$	1.45
	CoT-MATH	$2.67_{\pm 5.20}$	$2.67_{\pm 5.20}$	$2.67_{\pm 5.20}$	$2.67_{\pm 5.20}$	2.67

Table 23: Per-subset 13-gram overlap η (%; mean$_{\pm \text{std}}$ over non-members) for multi-subset domains. Domain-level values are macro-averaged across subsets; for DPO, they are averaged across the prompt, chosen, and rejected fields.

	Pre-training				Mid-training		Post-training			
	DCLM-Baseline	OpenWebMath	peS2o	StarCoder	Math Mix	Stack Exchange	SFT	DPO	RLVR	**Avg**
1B	0.49	0.52	0.48	0.50	0.52	0.48	0.51	0.47	0.51	0.50
7B	0.49	0.52	0.48	0.50	0.52	0.48	0.52	0.51	0.51	0.50
13B	0.49	0.52	0.48	0.50	0.52	0.48	0.52	0.49	0.51	0.50
32B	0.49	0.52	0.48	0.50	0.52	0.48	0.51	0.48	0.51	0.50
Avg	0.49	0.52	0.48	0.50	0.52	0.48	0.51	0.49	0.51	**0.50**

Table 24: Per-model-size blind detectability, measured by BoW classifier AUC. Multi-subset domains are macro-averaged across subsets, with per-subset AUCs reported in Table 25.

Analysis of Lexical Shift Across Domains. In OLMO-DETECT (SHIFTED), members are drawn at random rather than perturbed toward a target, so each domain shifts only along the axes on which its source corpus actually varies. Web pages, academic papers, and Q&A posts remain lexically uniform across quality tiers and publication years, so DCLM-Baseline, peS2o, and Stack Exchange shift on quality and temporal range instead: DCLM-Baseline's quality score rises from 0.018 to 0.131, peS2o's mean date moves back by nearly nine years, and Stack Exchange's question and answer votes roughly double, while their lexical values change by at most 0.03. By contrast, StarCoder, OpenWebMath, SFT, and DPO draw on corpora that vary lexically, through repository-specific identifiers, heterogeneous web pages, and multi-task multilingual data, and their lexical values rise two- to fourfold, from 0.09–0.15 to 0.32–0.36. In DPO, for instance, the dominant non-English script changes from Chinese to Cyrillic. Every domain is therefore shifted on at least one axis, and the shift is deliberately not maximized: **it emulates the unaligned construction of existing benchmarks rather than a worst case.**

H DETAILED DESCRIPTIONS OF EXPERIMENTAL SETTINGS

Implementation Details. We implement every compared method from its original public repository (Table 26). To ensure a fair comparison, every method is evaluated under an identical protocol with a batch size of 1 and a context length of 4,096 tokens for the generative methods that require text generation (Neigh., Guided, DCQ, CDD, and SelfCrit) and 1,024 tokens for the remaining non-generative methods. All experiments are run on NVIDIA A100 (80 GB) GPUs, using two GPUs for the 32B model and a single GPU for 1B to 13B models. Method-specific hyperparameters, where applicable, are tuned on the held-out development set; the search spaces and selected values are given in Table 27. For reproducibility, all experiments use a fixed random seed of 42.

Determinism and Statistical Reporting. We use a single random seed of 42 because the majority of our evaluated methods are deterministic. All non-generative unsupervised methods (PPL, Zlib, Lower, MinK, MinK++, DCPDD, ReCaLL, CAMIA, EMMIA) compute their statistics directly from token log-probabilities, without any sampling. Among the remaining methods: the two generative unsupervised methods, Guided and SelfCrit, decode greedily (`do_sample=False`, temperature

Domain	Subset	1B	7B	13B	32B	Avg
StarCoder	Python	0.48	0.48	0.48	0.48	0.48
	Java	0.51	0.51	0.51	0.51	0.51
Stack Exchange	Math Stack Exchange	0.47	0.47	0.47	0.47	0.47
	Stack Overflow	0.49	0.49	0.49	0.49	0.49
SFT	WildChat	0.52	0.54	0.54	0.52	0.53
	Aya	0.50	0.51	0.51	0.50	0.50
RLVR	CoT-GSM8K	0.51	0.51	0.51	0.51	0.51
	CoT-MATH	0.51	0.51	0.51	0.51	0.51

Table 25: Per-subset blind detectability for multi-subset domains, measured by BoW classifier AUC.

Method	Reference	Repository
PPL	Yeom et al. (2018)	— (standard baseline)
Zlib	Carlini et al. (2021)	— (standard baseline)
Lower	Carlini et al. (2021)	— (standard baseline)
MinK	Shi et al. (2023)	https://github.com/swj0419/detect-pretrain-code
MinK++	Zhang et al. (2024a)	https://github.com/zjysteven/mink-plus-plus
DCPDD	Zhang et al. (2024b)	https://github.com/zhang-wei-chao/DC-PDD
ReCaLL	Xie et al. (2024)	https://github.com/ruoyuxie/recall
CAMIA	Chang et al. (2025)	https://github.com/changhongyan123/context_aware_mia
PAC	Ye et al. (2024)	https://github.com/yyy01/PAC
EMMIA	Kim et al. (2026)	https://github.com/gyuwankim/em-mia
DCQ	Golchin & Surdeanu (2025)	https://github.com/shahriargolchin/DCQ
Neigh.	Mattern et al. (2023)	https://github.com/mireshghallah/neighborhood-curvature-mia
CDD	Dong et al. (2024)	https://github.com/YihongDong/CDD-TED4LLMs
Guided	Golchin & Surdeanu (2023)	https://github.com/shahriargolchin/time-travel-in-llms
SelfCrit	Tao et al. (2025)	https://github.com/yongding-tao/RL-Data-Contamination
FSD	Zhang et al. (2025)	https://github.com/ml-stat-Sustech/Fine-tuned-Score-Deviation
MIA-Tuner	Fu et al. (2025)	https://github.com/tsinghua-fib-lab/AAAI2025_MIA-Tuner
Sup-CAMIA	Chang et al. (2025)	https://github.com/changhongyan123/context_aware_mia

Table 26: Public repositories followed by our implementations of the compared MIAs. The final three rows are the supervised methods; all others are unsupervised.

0). SelfCrit's two responses are not sampled but produced from two different prompts, an initial answer and a follow-up requesting an alternative solution, so greedy decoding yields two distinct yet deterministic outputs. DCQ scores the target model with multiple-choice log-probabilities, which are likewise deterministic; its quiz options are generated once by an external `gpt-4o-mini` judge and cached on disk, shared across all model sizes and runs, so the identical quiz is reused on every run. Sup-CAMIA fits a convex logistic-regression classifier on deterministic features. For all of these methods, the seed has no effect on the reported scores.

The seed matters only for the methods that draw stochastic samples or train with random data ordering: Neigh. (neighbors sampled from a T5 mask-filling model), CDD (completions sampled from the target model at temperature 0.8), PAC (word-level random-swap augmentation), and the two supervised fine-tuning methods FSD (a LoRA adapter) and MIA-Tuner (a soft prompt). Due to the computational cost of repeated runs, none of these works reports results across multiple seeds in their papers; we follow this convention and fix both the framework seed (42) and each method's own internal seed parameter where one exists (e.g., PAC's and Neigh.'s), holding their random draws, augmentations, and training runs fixed across runs.

Since per-instance scores are fixed under a single seed, we quantify the uncertainty from the finite evaluation set with a nonparametric bootstrap. We resample members and non-members with replacement within every subset, recompute the full subset-to-domain-to-stage aggregation on each of $2{,}000$ resamples, and report the 2.5th and 97.5th percentiles as the 95% confidence interval, so the interval applies to the overall AUC rather than to any single evaluation. This captures uncertainty from the finite evaluation set but not seed-to-seed variance for the five stochastic methods above; we leave a full multi-seed analysis to future work. Intervals are given in Tables 28 and 29.

Method	Hyperparameter	Search space	Selected value			
			1B	7B	13B	32B
MinK	k	$\{0.05, \ldots, 1.0\}$	0.2			
MinK++	k	$\{0.05, \ldots, 1.0\}$	0.2			
DCPDD	α	$\{0.001, \ldots, 0.5\}$	0.001	0.2	0.05	0.2
ReCaLL	# prefix shots n	$\{1, \ldots, 12\}$	10	3	7	12
CAMIA	p-value composition	{Edgington, Fisher, George}	George	Fisher	George	George
PAC	(k_1, k_2)	$k_1 \in \{.05, .1, .2\},\ k_2 \in \{.1, .2, .3, .4\}$	(0.05, 0.10)			
DCQ	score type	{mean, min, κ, κ_{np}}P_{correct}	min	min	κ_{np}	κ
Neigh.	score type	$\{d, z\}$	d			
CDD	(α, L, r)	$\alpha \in \{0, \ldots, .3\}, L \in \{50, 100, 200\}, r \in \{.25, .5, .75\}$	$(.025, 50, .5)$	$(.075, 50, .5)$	$(.15, 50, .25)$	$(.3, 50, .25)$
Guided	score type	{ROUGE-guided, ROUGE-Δ}	ROUGE-Δ	ROUGE-Δ	ROUGE-Δ	ROUGE-guided

Table 27: Tunable hyperparameters, search spaces, and the values selected by development-set AUC on OLMO-DETECT.

Method	**AUC**	**95% CI**
PPL	0.65	[0.64, 0.67]
Zlib	0.68	[0.67, 0.70]
Lower	0.63	[0.62, 0.65]
MinK	0.66	[0.64, 0.67]
MinK++	0.56	[0.54, 0.57]
DCPDD	0.67	[0.65, 0.68]
PAC	0.66	[0.64, 0.67]
ReCaLL	0.46	[0.45, 0.47]
EMMIA	0.57	[0.56, 0.58]
CAMIA	0.65	[0.64, 0.67]
Neigh.	0.59	[0.58, 0.61]
Guided	0.49	[0.48, 0.50]
DCQ	0.49	[0.48, 0.51]
CDD	0.52	[0.52, 0.53]

Table 28: Bootstrap 95% confidence intervals for the overall AUC of each unsupervised MIA on OLMO-DETECT, i.e. the ***All*** column of Table 4 averaged over model sizes. Members and non-members are resampled with replacement within every subset and the full subset-to-domain-to-stage aggregation is recomputed on each of $2{,}000$ resamples, so the interval applies to the aggregate rather than to any single evaluation. SelfCrit is omitted because it covers only post-training. Point estimates match Table 4.

Continuous Scoring for Categorical Methods. DCQ (Golchin & Surdeanu, 2025) and Guided (Golchin & Surdeanu, 2023) originally produce categorical outputs that are not directly compatible with AUC computation. For DCQ, we replace the original single-letter output with the first-token logprobs over $\{A, B, C, D, E\}$ from a single forward pass on the model under evaluation. Each instance is presented through four bias compensator quiz (BCQ) prompts, with the original placed at slots $L \in \{A, B, C, D\}$ respectively; the per-slot score $P(L \mid \text{BCQ}_L)$ is the model's recognition probability for that placement. The final per-instance score is the mean (or minimum) of these four probabilities, with the choice tuned per (model, dataset) on the dev set. For Guided, we generate two greedy completions per instance: a *guided* completion (prompted with the source corpus name) and a *general* completion (no source cited). The per-instance score is the ROUGE-L F1 between the guided completion and the held-out second piece of the original paragraph, or its difference against the general completion; the variant is selected per (model, dataset) on the dev set. Both adaptations preserve the original methods' computations and use the same underlying signals (next-token logprobs for DCQ; ROUGE-L F1, the paper's own metric, for Guided); only the post-processing step that maps these signals to a categorical decision is replaced with continuous score extraction.

Excluded MIAs. Among existing MIAs that we have surveyed, four are excluded due to setting incompatibility: (i) methods that require access to a model other than $\mathcal{M}$, whether an existing smaller model or one separately fine-tuned, including Smaller Ref (Carlini et al., 2021) and SPV-MIA (Fu et al., 2024), since our setting assumes only grey-box access to $\mathcal{M}$, optionally extended to fine-tuning $\mathcal{M}$ itself, not the construction of an additional model; (ii) narrowly focused methods,

Method	Setting	AUC	95% CI
Sup-CAMIA	In-domain	0.68	[0.67, 0.68]
Sup-CAMIA	LODO	0.61	[0.60, 0.62]
MIA-Tuner	In-domain	0.55	[0.54, 0.56]
MIA-Tuner	LODO	0.49	[0.48, 0.50]

Table 29: Bootstrap 95% confidence intervals for the overall AUC of the supervised MIAs on OLMO-DETECT, i.e. the ***All*** column of Table 5 averaged over model sizes, computed as in Table 28. FSD is omitted because it is reported as a gain over its base detector rather than as an absolute AUC. Point estimates match Table 5.

Stage	Domain	Subsets	Split	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		Non-member	150	217,117	FastText: 0.01809 ± 0.00007	2018-01-19 ± 930d	0.08
			Member	150	217,295	FastText: 0.13129 ± 0.18279	2017-02-21 ± 1208d	
	peS2o		Non-member	268	216,981	—	2022-12-14 ± 11d	0.13
			Member	268	217,039	—	2014-02-03 ± 2657d	
	OpenWebMath		Non-member	151	217,183	—	2019-10-13 ± 853d	0.32
			Member	151	217,379	—	2011-10-01 ± 2936d	
	StarCoder	Python, Java	Non-member	156	217,156	GitHub Stars: 35.65 ± 66.66 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	0.36
			Member	156	217,284	GitHub Stars: 10.77 ± 17.94 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	
Mid-training	Stack Exchange	Stack Overflow, Math Stack Exchange	Non-member	430	217,208	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 1.99	2015-03-31 ± 1378d	0.08
			Member	430	217,209	Question Vote: 8.12 ± 7.08 Answer Vote: 10.02 ± 6.37	2013-05-31 ± 1213d	
Post-training	SFT†	WildChat, Aya	Non-member	225	217,057	—	—	0.34
			Member	225	217,238	—	—	
	DPO†	UF-WildChat	Non-member	113	217,045	—	—	0.32
			Member	113	217,154	—	—	

Table 30: Overall statistics of OLMO-DETECT (SHIFTED), spanning seven of the nine domains of OLMO-DETECT (GSM8K and RLVR are excluded, as neither has any shiftable axis). Non-members are unchanged from OLMO-DETECT, while members are randomly sampled subject only to matched instance and token counts, leaving Text **Quality**, **Temporal** Range, and **Lexical** Similarity unaligned. "—" denotes an axis that does not apply to a given domain, i.e., no quality thresholding is applied, or no timestamp metadata exists. **NM**: non-member; **M**: member. The **Subsets** column lists constituent subsets; blank entries indicate single-subset domains whose subset shares the domain's name. †SFT and DPO are constructed separately for each model size because their training data differ across sizes; the reported values correspond to the 1B model.

such as Name-cloze (Chang et al., 2023), since it requires text containing proper nouns to mask; (iii) methods with no open-source implementation, such as TS-Guessing (Deng et al., 2024).

I PER-DOMAIN AUC RESULTS FOR UNSUPERVISED MIAS ON OLMO-DETECT

This appendix provides detailed AUC results for unsupervised MIAs on OLMO-DETECT for each domain, complementing the stage-level aggregates in Table 4. Per-domain AUC results are shown in Tables 37–49.

J TPR@5% FPR RESULTS FOR UNSUPERVISED MIAS ON OLMO-DETECT

This appendix provides detailed per-domain TPR@5% FPR results for unsupervised MIAs on OLMO-DETECT. Stage-level aggregate results are reported in Table 50, while per-domain results are shown in Tables 51–58.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		NM	Dev	16	21,745	FastText: 0.01808 ± 0.00004	2018-01-18 ± 930.40d	0.08
				Test	134	195,372	FastText: 0.01809 ± 0.00008	2018-01-19 ± 930.19d	
				Total	**150**	**217,117**	FastText: 0.01809 ± 0.00007	2018-01-19 ± 930.21d	
			M	Dev	16	21,745	FastText: 0.13136 ± 0.19309	2017-02-21 ± 1204.63d	
				Test	134	195,550	FastText: 0.13128 ± 0.18152	2017-02-21 ± 1208.34d	
				Total	**150**	**217,295**	FastText: 0.13129 ± 0.18279	2017-02-21 ± 1207.94d	
	peS2o		NM	Dev	27	21,793	—	2022-12-12 ± 10.45d	0.13
				Test	241	195,188	—	2022-12-14 ± 11.29d	
				Total	**268**	**216,981**	—	2022-12-14 ± 11.22d	
			M	Dev	27	21,409	—	2014-02-03 ± 2656.90d	
				Test	241	195,630	—	2014-02-03 ± 2657.15d	
				Total	**268**	**217,039**	—	2014-02-03 ± 2657.12d	
	OpenWebMath		NM	Dev	15	21,761	—	2019-10-13 ± 850.83d	0.32
				Test	136	195,422	—	2019-10-13 ± 852.94d	
				Total	**151**	**217,183**	—	2019-10-13 ± 852.73d	
			M	Dev	15	21,820	—	2011-10-01 ± 2943.32d	
				Test	136	195,559	—	2011-10-02 ± 2935.43d	
				Total	**151**	**217,379**	—	2011-10-01 ± 2936.22d	
	StarCoder	Python, Java	NM	Dev	16	21,536	GitHub Stars: 36.19 ± 67.09 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.37 ± 0.04	—	0.36
				Test	140	195,620	GitHub Stars: 35.59 ± 66.61 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	
				Total	**156**	**217,156**	GitHub Stars: 35.65 ± 66.66 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.05	—	
			M	Dev	16	21,684	GitHub Stars: 10.69 ± 17.33 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.03	—	
				Test	140	195,600	GitHub Stars: 10.78 ± 18.00 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	
				Total	**156**	**217,284**	GitHub Stars: 10.77 ± 17.94 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	

Table 31: Detailed per-domain statistics for the pre-training stage of OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets; blank entries indicate single-subset domains whose subset shares the domain's name.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Stack Overflow, Math Stack Exchange	NM	Dev	43	21,760	Question Vote: 3.47 ± 2.20 Answer Vote: 5.12 ± 1.94	2015-03-27 ± 1378.74d	0.08
				Test	387	195,448	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 2.00	2015-03-31 ± 1378.12d	
				Total	**430**	**217,208**	Question Vote: 3.45 ± 2.19 Answer Vote: 5.12 ± 1.99	2015-03-31 ± 1378.18d	
			M	Dev	43	21,750	Question Vote: 8.16 ± 6.92 Answer Vote: 10.02 ± 6.41	2013-06-05 ± 1206.73d	
				Test	387	195,459	Question Vote: 8.12 ± 7.10 Answer Vote: 10.02 ± 6.37	2013-05-31 ± 1214.10d	
				Total	**430**	**217,209**	Question Vote: 8.12 ± 7.08 Answer Vote: 10.02 ± 6.37	2013-05-31 ± 1213.37d	

Table 32: Detailed per-domain statistics for the mid-training stage of OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets.

K PER-DOMAIN AUC RESULTS FOR SUPERVISED MIAS ON OLMO-DETECT

We report per-domain AUCs for the supervised MIAs in Tables 64–72, one table per domain, complementing the per-stage results of Table 5. Each table follows the two-part layout of Table 5: the top part gives the AUCs of Sup-CAMIA and MIA-Tuner under both settings, with Zlib repeated as an unsupervised reference, and the bottom part gives FSD's AUC gain over its base MIA, since FSD is only meaningful relative to the detector it is built on. **In-domain** trains on the evaluated domain's development set, and **LODO** trains on the other eight domains' development sets.

Two patterns hold across domains. First, supervision rarely pays: the better of Sup-CAMIA and MIA-Tuner exceeds Zlib on only five of the nine domains, ties on OpenWebMath, and falls below it

Stage	Domain	Subset	Model	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	WildChat, Aya	1B, 32B	NM	Dev	21	21,697	—	—	0.34
					Test	204	195,360	—	—	
					Total	**225**	**217,057**			
				M	Dev	21	21,727	—	—	
					Test	204	195,511	—	—	
					Total	**225**	**217,238**			
			7B, 13B	NM	Dev	16	21,959	—	—	0.34
					Test	148	195,165	—	—	
					Total	**164**	**217,124**			
				M	Dev	16	21,736	—	—	
					Test	148	195,437	—	—	
					Total	**164**	**217,173**			
	DPO	UF-WildChat	1B	NM	Dev	11	21,705	—	—	0.32
					Test	102	195,340	—	—	
					Total	**113**	**217,045**			
				M	Dev	11	21,554	—	—	
					Test	102	195,600	—	—	
					Total	**113**	**217,154**			
			7B	NM	Dev	10	21,704	—	—	0.28
					Test	100	195,163	—	—	
					Total	**110**	**216,867**			
				M	Dev	10	21,680	—	—	
					Test	100	195,345	—	—	
					Total	**110**	**217,025**			
			13B	NM	Dev	10	21,730	—	—	0.31
					Test	100	195,321	—	—	
					Total	**110**	**217,051**			
				M	Dev	10	21,720	—	—	
					Test	100	195,606	—	—	
					Total	**110**	**217,326**			
			32B	NM	Dev	10	21,838	—	—	0.40
					Test	100	195,333	—	—	
					Total	**110**	**217,171**			
				M	Dev	10	21,849	—	—	
					Test	100	195,320	—	—	
					Total	**110**	**217,169**			

Table 33: Detailed per-domain statistics for the post-training stage of OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set. The **Subset** column lists constituent subsets.

on DCLM-Baseline (-0.05), SFT (-0.03) and RLVR (-0.03); the one substantial gain is on DPO ($+0.10$), the domain where Zlib is weakest at 0.48. Second, fine-tuning on non-members seldom helps its base detector: of the 45 domain-by-base-MIA combinations, FSD improves on its base in 13 under In-domain and 11 under LODO, and its largest losses reach -0.10 or worse on DCLM-Baseline, where every base detector degrades under both settings.

L TPR@5%FPR RESULTS FOR SUPERVISED MIAS ON OLMO-DETECT

The main text reports supervised MIAs by AUC only, so we give their TPR@5% FPR here. Table 73 is the overall table, aggregated by training stage, and Tables 74–82 give one table per domain. All follow the two-part layout of Table 5: the top part reports Sup-CAMIA and MIA-Tuner under both settings, with Zlib repeated as an unsupervised reference, and the bottom part reports FSD's gain over the base MIA it is built on. **In-domain** trains on the evaluated domain's development set, and **LODO** trains on the other eight domains' development sets.

TPR@5% FPR reproduces the AUC picture and sharpens it. Overall, the best supervised configuration, Sup-CAMIA under In-domain, reaches 0.26 against Zlib's 0.28, so supervision does not improve the operating point that matters for an attacker; moving to LODO halves it to 0.14, and MIA-Tuner reaches only 0.09 In-domain and 0.05 under LODO, the latter indistinguishable from

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	StarCoder	Python	NM	Dev	10	19,364	GitHub Stars: 29.50 ± 61.41 Top-1 Freq: 0.31 ± 0.01 Top-2 Freq: 0.36 ± 0.04	—	0.37
				Test	90	175,594	GitHub Stars: 28.53 ± 59.41 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.38 ± 0.06	—	
				Total	**100**	**194,958**	GitHub Stars: 28.63 ± 59.62 Top-1 Freq: 0.30 ± 0.02 Top-2 Freq: 0.37 ± 0.05	—	
			M	Dev	10	19,295	GitHub Stars: 8.60 ± 17.62 Top-1 Freq: 0.08 ± 0.02 Top-2 Freq: 0.14 ± 0.03	—	
				Test	90	175,746	GitHub Stars: 9.02 ± 18.29 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	
				Total	**100**	**195,041**	GitHub Stars: 8.98 ± 18.23 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	
		Java	NM	Dev	6	2,172	GitHub Stars: 47.33 ± 74.29 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.38 ± 0.03	—	0.48
				Test	50	20,026	GitHub Stars: 48.28 ± 76.28 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.39 ± 0.03	—	
				Total	**56**	**22,198**	GitHub Stars: 48.18 ± 76.07 Top-1 Freq: 0.30 ± 0.01 Top-2 Freq: 0.39 ± 0.03	—	
			M	Dev	6	2,389	GitHub Stars: 14.17 ± 16.25 Top-1 Freq: 0.11 ± 0.03 Top-2 Freq: 0.17 ± 0.02	—	
				Test	50	19,854	GitHub Stars: 13.94 ± 17.02 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	
				Total	**56**	**22,243**	GitHub Stars: 13.96 ± 16.94 Top-1 Freq: 0.09 ± 0.03 Top-2 Freq: 0.15 ± 0.04	—	

Table 34: Detailed per-subset statistics for StarCoder in OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Stack Overflow	NM	Dev	37	18,489	Question Vote: 3.49 ± 2.26 Answer Vote: 5.14 ± 2.04	2015-02-06 ± 1381.39d	0.09
				Test	331	165,394	Question Vote: 3.48 ± 2.24 Answer Vote: 5.14 ± 2.05	2015-02-05 ± 1379.01d	
				Total	**368**	**183,883**	Question Vote: 3.48 ± 2.25 Answer Vote: 5.14 ± 2.05	2015-02-06 ± 1379.25d	
			M	Dev	37	18,488	Question Vote: 8.00 ± 6.70 Answer Vote: 10.00 ± 6.58	2013-07-30 ± 1269.85d	
				Test	331	165,394	Question Vote: 8.00 ± 6.98 Answer Vote: 10.00 ± 6.55	2013-07-25 ± 1281.27d	
				Total	**368**	**183,882**	Question Vote: 8.00 ± 6.95 Answer Vote: 10.00 ± 6.55	2013-07-26 ± 1280.13d	
		Math Stack Exchange	NM	Dev	6	3,271	Question Vote: 3.33 ± 1.80 Answer Vote: 5.00 ± 1.15	2016-01-22 ± 1323.12d	0.14
				Test	56	30,054	Question Vote: 3.25 ± 1.81 Answer Vote: 4.98 ± 1.64	2016-02-12 ± 1329.29d	
				Total	**62**	**33,325**	Question Vote: 3.26 ± 1.80 Answer Vote: 4.98 ± 1.60	2016-02-10 ± 1328.71d	
			M	Dev	6	3,262	Question Vote: 9.17 ± 8.07 Answer Vote: 10.17 ± 5.21	2012-06-28 ± 596.92d	
				Test	56	30,065	Question Vote: 8.84 ± 7.73 Answer Vote: 10.12 ± 5.19	2012-07-05 ± 596.75d	
				Total	**62**	**33,327**	Question Vote: 8.87 ± 7.76 Answer Vote: 10.13 ± 5.19	2012-07-04 ± 596.77d	

Table 35: Detailed per-subset statistics for Stack Exchange in OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

the 0.05 of a random detector. Per domain, the better of Sup-CAMIA and MIA-Tuner exceeds Zlib on four of the nine domains and falls below it on the other five, with the largest shortfall on StarCoder (-0.10) and the largest gains on Math Mix and DPO ($+0.08$ each). FSD again rarely helps

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Wildchat	NM	Dev	7	16,404	—	—	0.43
					Test	64	148,039	—	—	
					Total	**71**	**164,443**			
				M	Dev	7	16,388	—	—	
					Test	64	148,120	—	—	
					Total	**71**	**164,508**			
			Aya	NM	Dev	14	5,293	—	—	0.39
					Test	140	47,321	—	—	
					Total	**154**	**52,614**			
				M	Dev	14	5,339	—	—	
					Test	140	47,391	—	—	
					Total	**154**	**52,730**			
		7B, 13B	Wildchat	NM	Dev	7	16,715	—	—	0.41
					Test	66	148,305	—	—	
					Total	**73**	**165,020**			
				M	Dev	7	16,492	—	—	
					Test	66	148,578	—	—	
					Total	**73**	**165,070**			
			Aya	NM	Dev	9	5,244	—	—	0.43
					Test	82	46,860	—	—	
					Total	**91**	**52,104**			
				M	Dev	9	5,244	—	—	
					Test	82	46,859	—	—	
					Total	**91**	**52,103**			

Table 36: Detailed per-subset statistics for SFT in OLMO-DETECT (SHIFTED). **NM**: non-member; **M**: member. **Dev**: development set; **Test**: test set.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.60	0.59	0.60	0.57	0.46	0.57	0.61	0.46	0.52	0.60	**0.64**	0.54	0.50	0.50	0.55
7B	0.64	0.64	0.62	0.61	0.53	0.64	0.65	0.50	0.54	**0.65**	0.61	0.49	0.43	0.49	0.57
13B	0.65	0.65	0.63	0.63	0.52	0.64	**0.68**	0.62	0.53	0.66	0.62	0.47	0.47	0.51	0.59
32B	0.66	0.67	0.66	0.64	0.61	0.65	**0.69**	0.53	0.50	0.64	0.61	0.47	0.64	0.51	0.61
Model-Avg	0.64	0.64	0.63	0.61	0.53	0.62	**0.66**	0.53	0.52	0.64	0.62	0.49	0.51	0.50	0.58

Table 37: Comparison of unsupervised MIAs on the DCLM-Baseline domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

its base MIA: of the 45 domain-by-base-MIA combinations it improves in 12 under In-domain and 11 under LODO.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.47	0.49	0.48	0.50	0.51	0.52	0.48	0.51	**0.63**	0.50	0.49	0.44	0.49	0.50	0.50
7B	0.47	0.49	0.48	0.48	0.49	0.51	0.47	0.55	**0.70**	0.47	0.51	0.52	0.52	0.50	0.51
13B	0.48	0.50	0.50	0.49	0.50	0.51	0.51	0.51	**0.70**	0.47	0.51	0.51	0.54	0.50	0.52
32B	0.49	0.51	0.50	0.51	0.48	0.52	0.51	0.52	**0.71**	0.47	0.53	0.44	0.64	0.49	0.52
Model-Avg	0.48	0.50	0.49	0.50	0.50	0.51	0.49	0.52	**0.69**	0.48	0.51	0.48	0.55	0.50	0.51

Table 38: Comparison of unsupervised MIAs on the peS2o domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.58	0.65	0.60	0.60	0.49	0.65	**0.69**	0.60	0.53	0.55	0.53	0.43	0.59	0.52	0.57
7B	0.60	0.65	0.62	0.61	0.50	0.67	**0.68**	0.60	0.53	0.54	0.53	0.55	0.46	0.54	0.58
13B	0.60	0.64	0.59	0.60	0.50	0.68	**0.70**	0.60	0.52	0.60	0.52	0.45	0.54	0.56	0.58
32B	0.60	0.63	0.58	0.62	0.48	0.69	**0.70**	0.56	0.64	0.61	0.52	0.58	0.45	0.54	0.59
Model-Avg	0.59	0.64	0.60	0.61	0.49	0.67	**0.69**	0.59	0.56	0.58	0.53	0.50	0.51	0.54	0.58

Table 39: Comparison of unsupervised MIAs on the StarCoder domain of OLMO-DETECT, evaluated via AUC (macro-averaged over its subsets). Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.51	0.56	0.61	0.55	0.48	**0.61**	0.54	0.44	0.54	0.56	0.54	0.47	0.52	0.51	0.53
7B	0.51	0.57	**0.59**	0.55	0.50	0.57	0.56	0.49	0.53	0.46	0.52	0.42	0.24	0.48	0.50
13B	0.52	0.58	0.61	0.56	0.53	**0.63**	0.58	0.48	0.52	0.45	0.55	0.44	0.42	0.49	0.53
32B	0.57	0.64	0.59	0.60	0.59	**0.65**	0.60	0.44	0.60	0.43	0.55	0.43	0.51	0.48	0.55
Model-Avg	0.53	0.59	0.60	0.56	0.52	**0.61**	0.57	0.46	0.55	0.47	0.54	0.44	0.42	0.49	0.53

Table 40: Comparison of unsupervised MIAs on the OpenWebMath domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.73	0.76	0.72	**0.76**	0.57	0.69	0.72	0.32	0.50	0.75	0.64	0.49	0.46	0.52	0.62
7B	0.85	0.89	0.83	0.86	0.52	**0.91**	0.83	0.05	0.51	0.88	0.72	0.50	0.51	0.55	0.67
13B	0.98	**0.99**	0.93	0.99	0.58	0.99	0.96	0.01	0.48	0.99	0.89	0.48	0.53	0.68	0.75
32B	0.98	0.99	0.93	0.97	0.51	0.98	0.92	0.02	0.50	**0.99**	0.91	0.55	0.51	0.74	0.75
Model-Avg	0.89	**0.91**	0.85	0.89	0.55	0.89	0.86	0.10	0.50	0.90	0.79	0.51	0.50	0.62	0.70

Table 41: Comparison of unsupervised MIAs on the GSM8K subset of OLMO-DETECT's Math Mix domain, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.55	**0.59**	0.50	0.54	0.49	0.53	0.59	0.54	0.56	0.58	0.55	0.54	0.54	0.51	0.54
7B	0.58	**0.64**	0.55	0.60	0.56	0.61	0.63	0.53	0.57	0.59	0.56	0.45	0.40	0.53	0.56
13B	0.60	**0.67**	0.56	0.62	0.61	0.67	0.66	0.66	0.65	0.61	0.58	0.49	0.64	0.52	0.61
32B	0.63	0.70	0.53	0.65	0.61	0.68	0.69	**0.73**	0.69	0.64	0.60	0.54	0.64	0.50	0.63
Model-Avg	0.59	**0.65**	0.53	0.60	0.57	0.62	0.64	0.61	0.62	0.61	0.57	0.51	0.55	0.51	0.59

Table 42: Comparison of unsupervised MIAs on the Stack Exchange domain of OLMO-DETECT, evaluated via AUC (macro-averaged over its subsets). Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.62	0.64	0.49	0.63	**0.73**	0.56	0.51	0.61	0.57	0.49	0.54	0.48	0.51	0.51	0.56
7B	0.73	0.70	0.55	0.71	**0.77**	0.70	0.59	0.74	0.59	0.74	0.60	0.51	0.43	0.50	0.63
13B	0.65	0.61	0.50	0.64	**0.70**	0.60	0.55	0.65	0.55	0.57	0.51	0.54	0.44	0.51	0.57
32B	0.64	0.64	0.50	0.65	**0.69**	0.60	0.51	0.69	0.55	0.54	0.52	0.44	0.45	0.50	0.57
Model-Avg	0.66	0.65	0.51	0.66	**0.72**	0.61	0.54	0.67	0.56	0.58	0.54	0.49	0.46	0.51	0.58

Table 43: Comparison of unsupervised MIAs on the SFT domain of OLMO-DETECT, evaluated via AUC (macro-averaged over its subsets). Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
1B	0.44	0.55	**0.64**	0.44	0.45	0.47	0.62	0.54	0.60	0.59	0.54	0.53	0.45	0.53	0.43	0.52
7B	0.41	0.39	0.50	0.40	0.40	0.40	0.43	0.48	0.50	**0.60**	0.46	0.48	0.30	0.53	0.53	0.45
13B	0.38	0.47	**0.64**	0.38	0.45	0.39	0.53	0.59	0.49	0.57	0.49	0.45	0.50	0.53	0.43	0.49
32B	0.46	0.53	0.62	0.48	0.47	0.48	0.62	**0.65**	0.39	0.55	0.60	0.47	0.47	0.48	0.55	0.52
Model-Avg	0.42	0.48	**0.60**	0.43	0.44	0.43	0.55	0.57	0.49	0.58	0.52	0.48	0.43	0.52	0.48	0.50

Table 44: Comparison of unsupervised MIAs on the UF-Wildchat subset of OLMO-DETECT's DPO domain, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
1B	0.77	0.79	0.64	**0.79**	0.60	0.74	0.69	0.36	0.66	0.68	0.51	0.50	0.43	0.48	0.50	0.61
7B	0.89	0.88	0.80	**0.90**	0.69	0.87	0.77	0.25	0.73	0.82	0.60	0.53	0.48	0.48	0.48	0.68
13B	0.95	0.94	0.81	**0.95**	0.62	0.92	0.75	0.18	0.76	0.85	0.58	0.47	0.50	0.46	0.46	0.68
32B	0.94	0.91	0.90	**0.94**	0.72	0.92	0.77	0.28	0.47	0.89	0.60	0.47	0.50	0.49	0.41	0.68
Model-Avg	0.89	0.88	0.79	**0.90**	0.66	0.86	0.75	0.27	0.65	0.81	0.57	0.49	0.48	0.48	0.46	0.66

Table 45: Comparison of unsupervised MIAs on the RLVR domain of OLMO-DETECT, evaluated via AUC (macro-averaged over its subsets). Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Python	1B	0.54	0.61	0.58	0.55	0.50	0.61	**0.69**	0.52	0.49	0.58	0.52	0.38	0.57	0.52	0.55
	7B	0.55	0.62	0.60	0.56	0.53	0.61	**0.68**	0.51	0.50	0.58	0.52	0.48	0.51	0.51	0.56
	13B	0.55	0.62	0.58	0.56	0.52	0.62	**0.67**	0.56	0.51	0.62	0.51	0.48	0.54	0.52	0.56
	32B	0.54	0.60	0.55	0.55	0.53	0.61	**0.66**	0.53	0.62	0.60	0.53	0.56	0.49	0.56	0.57
	Model-Avg	0.54	0.62	0.58	0.55	0.52	0.61	**0.68**	0.53	0.53	0.60	0.52	0.48	0.53	0.53	0.56
Java	1B	0.62	0.68	0.61	0.65	0.47	0.69	**0.70**	0.68	0.57	0.51	0.54	0.47	0.61	0.51	0.59
	7B	0.64	0.68	0.63	0.65	0.48	**0.73**	0.69	0.68	0.56	0.50	0.55	0.63	0.41	0.57	0.60
	13B	0.64	0.65	0.61	0.65	0.47	**0.74**	0.72	0.64	0.53	0.58	0.54	0.43	0.54	0.60	0.60
	32B	0.66	0.66	0.62	0.69	0.43	**0.77**	0.73	0.60	0.66	0.62	0.51	0.60	0.40	0.51	0.60
	Model-Avg	0.64	0.67	0.62	0.66	0.46	**0.73**	0.71	0.65	0.58	0.55	0.53	0.53	0.49	0.55	0.60

Table 46: Comparison of unsupervised MIAs on the subsets of the StarCoder domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Math Stack Exchange	1B	0.55	0.60	0.49	0.54	0.49	0.55	**0.62**	0.57	0.58	0.60	0.57	0.60	0.59	0.52	0.56
	7B	0.57	0.63	0.53	0.58	0.55	0.59	**0.67**	0.51	0.62	0.61	0.57	0.41	0.36	0.54	0.55
	13B	0.60	0.67	0.55	0.62	0.62	0.68	**0.71**	0.63	0.69	0.64	0.60	0.45	0.63	0.51	0.61
	32B	0.62	0.71	0.51	0.65	0.61	0.70	0.71	**0.79**	0.74	0.66	0.61	0.57	0.59	0.50	0.64
	Model-Avg	0.59	0.65	0.52	0.60	0.57	0.63	**0.68**	0.63	0.66	0.63	0.59	0.51	0.54	0.52	0.59
Stack Overflow	1B	0.56	**0.58**	0.50	0.54	0.49	0.52	0.55	0.51	0.55	0.57	0.54	0.49	0.49	0.49	0.53
	7B	0.59	**0.64**	0.56	0.61	0.57	0.63	0.60	0.54	0.51	0.56	0.55	0.49	0.43	0.51	0.56
	13B	0.60	0.67	0.56	0.63	0.60	0.66	0.61	**0.70**	0.60	0.59	0.57	0.52	0.65	0.52	0.60
	32B	0.63	**0.70**	0.55	0.65	0.61	0.67	0.66	0.67	0.63	0.63	0.58	0.52	0.68	0.49	0.62
	Model-Avg	0.60	**0.65**	0.55	0.61	0.57	0.62	0.60	0.60	0.57	0.59	0.56	0.50	0.56	0.50	0.58

Table 47: Comparison of unsupervised MIAs on the subsets of the Stack Exchange domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Aya	1B	0.56	0.58	0.50	0.54	**0.67**	0.54	0.50	0.48	0.58	0.54	0.54	0.49	0.48	0.50	0.53
	7B	**0.77**	0.70	0.53	0.72	0.76	0.72	0.67	0.69	0.59	0.76	0.69	0.52	0.47	0.53	0.65
	13B	**0.69**	0.60	0.49	0.66	0.64	0.62	0.60	0.56	0.51	0.67	0.59	0.51	0.49	0.55	0.59
	32B	0.55	0.56	0.49	0.54	0.59	0.55	0.49	**0.60**	0.44	0.50	0.51	0.38	0.51	0.51	0.52
	Model-Avg	0.64	0.61	0.51	0.61	**0.66**	0.61	0.57	0.58	0.53	0.62	0.58	0.47	0.49	0.52	0.57
WildChat	1B	0.69	0.70	0.48	0.71	**0.79**	0.58	0.51	0.75	0.55	0.44	0.54	0.47	0.54	0.52	0.59
	7B	0.70	0.71	0.57	0.70	**0.79**	0.69	0.52	0.79	0.59	0.73	0.50	0.49	0.39	0.48	0.62
	13B	0.62	0.62	0.51	0.62	**0.76**	0.57	0.50	0.74	0.60	0.46	0.43	0.57	0.39	0.47	0.56
	32B	0.72	0.72	0.51	0.76	**0.79**	0.65	0.52	0.77	0.66	0.57	0.53	0.51	0.38	0.50	0.61
	Model-Avg	0.68	0.69	0.52	0.70	**0.78**	0.62	0.51	0.76	0.60	0.55	0.50	0.51	0.43	0.49	0.60

Table 48: Comparison of unsupervised MIAs on the subsets of the SFT domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
CoT-GSM8K	1B	0.85	**0.89**	0.73	0.88	0.62	0.80	0.78	0.21	0.80	0.75	0.49	0.51	0.42	0.51	0.47	0.65
	7B	0.98	0.99	0.87	**0.99**	0.78	0.98	0.89	0.10	0.97	0.98	0.62	0.51	0.50	0.53	0.50	0.75
	13B	**1.00**	1.00	0.90	1.00	0.64	0.99	0.78	0.01	1.00	0.98	0.56	0.48	0.41	0.49	0.39	0.71
	32B	**1.00**	1.00	0.95	1.00	0.79	1.00	0.87	0.04	0.50	0.99	0.61	0.49	0.51	0.50	0.35	0.71
	Model-Avg	0.96	**0.97**	0.86	0.97	0.71	0.94	0.83	0.09	0.82	0.93	0.57	0.50	0.46	0.51	0.43	0.70
CoT-MATH	1B	0.70	0.68	0.54	**0.71**	0.58	0.69	0.60	0.51	0.52	0.61	0.52	0.50	0.43	0.46	0.53	0.57
	7B	0.80	0.77	0.74	**0.81**	0.61	0.75	0.66	0.41	0.49	0.66	0.58	0.54	0.46	0.43	0.46	0.61
	13B	0.90	0.87	0.72	**0.91**	0.60	0.86	0.73	0.36	0.51	0.71	0.60	0.46	0.58	0.44	0.52	0.65
	32B	0.87	0.82	0.86	**0.88**	0.65	0.85	0.68	0.52	0.43	0.79	0.58	0.44	0.50	0.48	0.47	0.65
	Model-Avg	0.82	0.79	0.72	**0.83**	0.61	0.79	0.67	0.45	0.49	0.69	0.57	0.49	0.49	0.45	0.50	0.62

Table 49: Comparison of unsupervised MIAs on the subsets of the RLVR domain of OLMO-DETECT, evaluated via AUC. Methods are grouped by family as in Table 4. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Model	Stage	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
1B	Pre	0.07	0.09	0.08	0.09	0.07	**0.12**	0.12	0.07	0.10	0.09	0.05	0.03	0.07	0.06	—	0.08
	Mid	0.14	**0.17**	0.15	0.15	0.09	0.10	0.14	0.07	0.09	0.15	0.10	0.06	0.07	0.06	—	0.11
	Post	**0.20**	0.18	0.12	0.20	0.15	0.12	0.13	0.10	0.09	0.17	0.05	0.05	0.04	0.07	0.07	0.12
	All	0.13	**0.15**	0.12	0.15	0.10	0.11	0.13	0.08	0.09	0.14	0.07	0.05	0.06	0.06	0.07	0.10
7B	Pre	0.08	0.13	0.10	0.11	0.09	0.15	0.13	0.08	**0.18**	0.12	0.06	0.05	0.05	0.06	—	0.10
	Mid	0.24	0.33	0.15	0.28	0.10	**0.39**	0.24	0.02	0.07	0.27	0.16	0.05	0.05	0.07	—	0.17
	Post	**0.36**	0.32	0.18	0.34	0.23	0.30	0.15	0.09	0.22	0.27	0.07	0.05	0.03	0.07	0.04	0.18
	All	0.23	0.26	0.14	0.24	0.14	**0.28**	0.17	0.06	0.16	0.22	0.10	0.05	0.04	0.06	0.04	0.15
13B	Pre	0.10	0.13	0.09	0.11	0.07	0.13	0.17	0.06	**0.19**	0.10	0.08	0.04	0.05	0.06	—	0.10
	Mid	0.49	**0.57**	0.33	0.51	0.14	0.55	0.49	0.05	0.07	0.50	0.31	0.06	0.05	0.11	—	0.30
	Post	**0.34**	0.34	0.26	0.33	0.12	0.28	0.14	0.15	0.22	0.28	0.07	0.05	0.06	0.08	0.06	0.19
	All	0.31	**0.35**	0.22	0.31	0.11	0.32	0.26	0.09	0.16	0.30	0.15	0.05	0.05	0.08	0.06	0.19
32B	Pre	0.12	0.16	0.07	0.12	0.10	0.16	**0.20**	0.06	0.20	0.12	0.08	0.04	0.07	0.06	—	0.11
	Mid	0.49	**0.58**	0.39	0.49	0.13	0.56	0.46	0.08	0.12	0.52	0.36	0.11	0.10	0.14	—	0.32
	Post	0.35	0.31	0.19	**0.37**	0.18	0.32	0.17	0.13	0.04	0.31	0.09	0.04	0.05	0.07	0.07	0.18
	All	0.32	**0.35**	0.22	0.32	0.14	0.35	0.28	0.09	0.12	0.31	0.18	0.06	0.07	0.09	0.07	0.20
Model-Avg	Pre	0.09	0.13	0.09	0.10	0.08	0.14	0.15	0.07	**0.17**	0.11	0.07	0.04	0.06	0.06	—	0.10
	Mid	0.34	**0.41**	0.26	0.36	0.11	0.40	0.33	0.05	0.09	0.36	0.24	0.07	0.07	0.10	—	0.23
	Post	**0.31**	0.29	0.19	0.31	0.17	0.25	0.15	0.12	0.14	0.26	0.07	0.05	0.04	0.07	0.06	0.17
	All	0.25	**0.28**	0.18	0.26	0.12	0.26	0.21	0.08	0.13	0.24	0.12	0.05	0.06	0.07	0.06	0.16

Table 50: Comparison of unsupervised MIAs on OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family: likelihood-based, perturbation-based (**Pert.**), prompt-based and output-distribution-based (**Out.-Dist.**). **Pre/Mid/Post** report macro-averaged TPRs over the domains of each stage, while **All** macro-averages the three stages. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR and thus has only **Post** and **All** values, averaged over these two domains. Per-domain results are provided in Tables 51–63.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.09	0.06	0.14	0.13	0.01	0.12	**0.16**	0.03	0.06	0.13	0.09	0.04	0.03	0.06	0.08
7B	0.15	0.11	0.16	0.21	0.06	0.18	0.20	0.03	0.07	**0.22**	0.08	0.06	0.04	0.04	0.12
13B	0.16	0.13	0.14	0.20	0.05	0.15	**0.26**	0.06	0.08	0.23	0.10	0.04	0.03	0.06	0.12
32B	0.21	0.18	0.13	0.22	0.11	0.18	**0.34**	0.06	0.12	0.26	0.11	0.04	0.07	0.07	0.15
Model-Avg	0.15	0.12	0.14	0.19	0.06	0.16	**0.24**	0.04	0.08	0.21	0.10	0.05	0.04	0.06	0.12

Table 51: Comparison of unsupervised MIAs on the DCLM-Baseline domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.06	0.07	0.05	0.09	0.07	0.07	0.06	0.05	**0.18**	0.06	0.04	0.05	0.04	0.05	0.07
7B	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.05	**0.47**	0.09	0.05	0.03	0.10	0.05	0.09
13B	0.07	0.07	0.08	0.07	0.05	0.09	0.10	0.05	**0.44**	0.05	0.07	0.05	0.07	0.05	0.09
32B	0.08	0.09	0.08	0.10	0.08	0.10	0.09	0.07	**0.48**	0.05	0.06	0.05	0.11	0.04	0.10
Model-Avg	0.07	0.07	0.07	0.08	0.07	0.08	0.08	0.05	**0.39**	0.06	0.05	0.05	0.08	0.05	0.09

Table 52: Comparison of unsupervised MIAs on the peS2o domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.08	0.17	0.07	0.08	0.13	0.16	0.17	**0.17**	0.09	0.11	0.05	0.01	0.16	0.07	0.11
7B	0.08	**0.27**	0.10	0.08	0.08	0.23	0.22	0.22	0.10	0.11	0.04	0.05	0.06	0.12	0.13
13B	0.09	0.24	0.08	0.09	0.07	0.20	**0.27**	0.09	0.11	0.10	0.07	0.03	0.01	0.07	0.11
32B	0.08	0.19	0.04	0.08	0.03	0.24	**0.26**	0.08	0.10	0.12	0.06	0.04	0.06	0.10	0.11
Model-Avg	0.08	0.22	0.07	0.08	0.08	0.21	**0.23**	0.14	0.10	0.11	0.06	0.03	0.07	0.09	0.11

Table 53: Comparison of unsupervised MIAs on the StarCoder domain of OLMO-DETECT, evaluated via TPR@5% FPR (macro-averaged over its subsets). Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.04	0.07	0.08	0.06	0.08	**0.12**	0.07	0.01	0.09	0.07	0.04	0.03	0.05	0.07	0.06
7B	0.04	0.07	0.07	0.06	**0.14**	0.12	0.02	0.01	0.09	0.06	0.07	0.04	0.00	0.04	0.06
13B	0.06	0.10	0.04	0.06	0.11	0.10	0.04	0.04	**0.15**	0.04	0.07	0.03	0.08	0.03	0.07
32B	0.10	**0.17**	0.04	0.07	0.15	0.13	0.11	0.03	0.10	0.04	0.07	0.02	0.04	0.02	0.08
Model-Avg	0.06	0.10	0.06	0.06	**0.12**	0.12	0.06	0.03	0.11	0.05	0.06	0.03	0.04	0.04	0.07

Table 54: Comparison of unsupervised MIAs on the OpenWebMath domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.21	**0.25**	0.17	0.25	0.14	0.12	0.16	0.03	0.04	0.22	0.11	0.04	0.04	0.07	0.13
7B	0.43	0.54	0.21	0.48	0.13	**0.60**	0.37	0.00	0.05	0.49	0.22	0.05	0.06	0.07	0.26
13B	0.92	**0.97**	0.54	0.93	0.21	0.94	0.78	0.00	0.03	0.95	0.53	0.05	0.04	0.17	0.50
32B	0.89	0.96	0.61	0.84	0.14	0.92	0.67	0.00	0.06	**0.96**	0.63	0.12	0.06	0.24	0.51
Model-Avg	0.61	**0.68**	0.38	0.63	0.16	0.65	0.49	0.01	0.05	0.66	0.37	0.06	0.05	0.14	0.35

Table 55: Comparison of unsupervised MIAs on the GSM8K subset of OLMO-DETECT's Math Mix domain, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.06	0.08	0.12	0.05	0.03	0.08	0.12	0.10	**0.13**	0.07	0.10	0.07	0.10	0.06	0.08
7B	0.06	0.12	0.09	0.08	0.06	**0.18**	0.12	0.03	0.09	0.06	0.10	0.05	0.05	0.06	0.08
13B	0.07	0.17	0.13	0.08	0.07	0.16	**0.19**	0.10	0.10	0.05	0.10	0.07	0.06	0.05	0.10
32B	0.09	0.20	0.17	0.13	0.12	0.21	**0.25**	0.16	0.17	0.07	0.10	0.10	0.14	0.05	0.14
Model-Avg	0.07	0.14	0.13	0.08	0.07	0.16	**0.17**	0.10	0.12	0.06	0.10	0.07	0.09	0.05	0.10

Table 56: Comparison of unsupervised MIAs on the Stack Exchange domain of OLMO-DETECT, evaluated via TPR@5% FPR (macro-averaged over its subsets). Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
1B	0.22	0.14	0.09	0.21	**0.33**	0.12	0.08	0.12	0.04	0.15	0.07	0.05	0.05	0.05	0.12
7B	0.40	0.29	0.13	0.36	**0.50**	0.28	0.17	0.17	0.07	0.22	0.06	0.06	0.02	0.06	0.20
13B	0.19	0.14	0.07	0.18	**0.26**	0.10	0.06	0.11	0.06	0.14	0.09	0.10	0.07	0.08	0.12
32B	0.28	0.15	0.07	0.32	**0.33**	0.19	0.05	0.17	0.07	0.21	0.07	0.08	0.00	0.08	0.15
Model-Avg	0.27	0.18	0.09	0.27	**0.36**	0.17	0.09	0.14	0.06	0.18	0.07	0.07	0.04	0.07	0.15

Table 57: Comparison of unsupervised MIAs on the SFT domain of OLMO-DETECT, evaluated via TPR@5% FPR (macro-averaged over its subsets). Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
1B	0.07	0.06	0.14	0.08	0.04	0.07	0.12	0.16	0.08	**0.22**	0.04	0.04	0.03	0.10	0.12	0.09
7B	0.02	0.03	0.07	0.01	0.02	0.01	0.01	0.07	0.09	0.07	0.04	0.05	0.00	**0.10**	0.04	0.04
13B	0.03	0.06	0.28	0.04	0.04	0.03	0.06	**0.33**	0.10	0.08	0.03	0.03	0.06	0.11	0.06	0.09
32B	0.04	0.09	0.09	0.06	0.07	0.04	0.13	**0.19**	0.01	0.06	0.03	0.03	0.07	0.07	0.11	0.07
Model-Avg	0.04	0.06	0.15	0.05	0.04	0.04	0.08	**0.19**	0.07	0.11	0.03	0.04	0.04	0.09	0.08	0.07

Table 58: Comparison of unsupervised MIAs on the UF-Wildchat subset of OLMO-DETECT's DPO domain, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
	PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
1B	0.29	**0.35**	0.12	0.33	0.08	0.16	0.19	0.02	0.15	0.14	0.03	0.06	0.04	0.04	0.02	0.14
7B	**0.66**	0.65	0.33	0.66	0.16	0.60	0.25	0.02	0.51	0.53	0.11	0.04	0.05	0.03	0.05	0.31
13B	**0.82**	0.81	0.41	0.78	0.06	0.71	0.30	0.00	0.52	0.62	0.11	0.02	0.05	0.05	0.05	0.36
32B	0.71	0.68	0.43	**0.74**	0.15	0.71	0.34	0.03	0.04	0.67	0.17	0.03	0.06	0.05	0.02	0.32
Model-Avg	0.62	0.62	0.32	**0.63**	0.11	0.54	0.27	0.02	0.30	0.49	0.10	0.04	0.05	0.05	0.04	0.28

Table 59: Comparison of unsupervised MIAs on the RLVR domain of OLMO-DETECT, evaluated via TPR@5% FPR (macro-averaged over its subsets). Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Python	1B	0.10	**0.23**	0.10	0.10	0.12	0.20	0.16	0.09	0.06	0.17	0.06	0.00	0.14	0.08	0.11
	7B	0.10	0.22	0.12	0.10	0.12	0.19	**0.23**	0.09	0.08	0.13	0.06	0.07	0.05	0.02	0.11
	13B	0.11	0.22	0.10	0.12	0.07	0.18	**0.26**	0.06	0.08	0.16	0.04	0.06	0.01	0.07	0.11
	32B	0.10	0.14	0.03	0.10	0.07	0.17	**0.22**	0.09	0.10	0.16	0.04	0.04	0.06	0.13	0.10
	Model-Avg	0.10	0.21	0.09	0.10	0.09	0.18	**0.22**	0.08	0.08	0.15	0.05	0.04	0.06	0.07	0.11
Java	1B	0.06	0.10	0.04	0.06	0.14	0.12	0.18	**0.26**	0.12	0.06	0.04	0.02	0.18	0.07	0.10
	7B	0.06	0.32	0.08	0.06	0.04	0.28	0.20	**0.36**	0.12	0.08	0.02	0.04	0.08	0.22	0.14
	13B	0.06	0.26	0.06	0.06	0.08	0.22	**0.28**	0.12	0.14	0.04	0.10	0.00	0.00	0.08	0.11
	32B	0.06	0.24	0.04	0.06	0.00	**0.32**	0.30	0.08	0.10	0.08	0.08	0.04	0.06	0.08	0.11
	Model-Avg	0.06	0.23	0.06	0.06	0.07	0.23	**0.24**	0.20	0.12	0.07	0.06	0.03	0.08	0.11	0.12

Table 60: Comparison of unsupervised MIAs on the subsets of the StarCoder domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Math Stack Exchange	1B	0.07	0.07	0.16	0.04	0.04	0.11	**0.20**	0.14	0.20	0.11	0.11	0.09	0.16	0.08	0.11
	7B	0.05	0.07	0.07	0.07	0.02	**0.25**	0.16	0.04	0.14	0.08	0.11	0.02	0.02	0.05	0.08
	13B	0.07	0.16	0.15	0.07	0.04	0.20	**0.27**	0.11	0.11	0.02	0.12	0.09	0.07	0.04	0.11
	32B	0.11	0.25	0.21	0.14	0.11	0.25	**0.39**	0.23	0.20	0.05	0.11	0.11	0.12	0.05	0.17
	Model-Avg	0.08	0.14	0.15	0.08	0.05	0.20	**0.25**	0.13	0.16	0.06	0.11	0.08	0.09	0.06	0.12
Stack Overflow	1B	0.06	**0.10**	0.09	0.07	0.03	0.05	0.05	0.06	0.06	0.04	0.08	0.05	0.05	0.03	0.06
	7B	0.06	**0.16**	0.11	0.09	0.10	0.11	0.08	0.03	0.05	0.05	0.08	0.08	0.08	0.07	0.08
	13B	0.06	**0.18**	0.11	0.08	0.10	0.13	0.11	0.10	0.10	0.08	0.08	0.05	0.06	0.07	0.09
	32B	0.08	0.16	0.12	0.11	0.13	**0.17**	0.11	0.09	0.15	0.08	0.10	0.09	0.15	0.04	0.11
	Model-Avg	0.06	**0.15**	0.10	0.09	0.09	0.11	0.09	0.07	0.09	0.06	0.09	0.07	0.08	0.05	0.09

Table 61: Comparison of unsupervised MIAs on the subsets of the Stack Exchange domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	
Aya	1B	0.12	0.02	0.16	0.16	**0.24**	0.09	0.04	0.09	0.06	0.14	0.08	0.06	0.04	0.05	0.10
	7B	**0.48**	0.27	0.12	0.40	0.35	0.26	0.20	0.18	0.09	0.41	0.07	0.10	0.05	0.11	0.22
	13B	**0.28**	0.15	0.06	0.28	0.23	0.15	0.10	0.09	0.04	0.26	0.09	0.13	0.11	0.13	0.15
	32B	0.17	0.04	0.11	0.14	0.10	0.06	0.04	**0.22**	0.07	0.08	0.10	0.04	0.01	0.14	0.09
	Model-Avg	**0.26**	0.12	0.11	0.25	0.23	0.14	0.09	0.15	0.06	0.22	0.08	0.08	0.05	0.11	0.14
WildChat	1B	0.31	0.25	0.02	0.26	**0.42**	0.14	0.12	0.14	0.03	0.16	0.06	0.05	0.05	0.05	0.15
	7B	0.32	0.32	0.14	0.32	**0.64**	0.30	0.15	0.17	0.05	0.03	0.05	0.02	0.00	0.02	0.18
	13B	0.09	0.14	0.07	0.08	**0.29**	0.05	0.03	0.14	0.08	0.03	0.09	0.08	0.03	0.03	0.09
	32B	0.39	0.25	0.03	0.49	**0.56**	0.33	0.06	0.11	0.08	0.34	0.05	0.11	0.00	0.02	0.20
	Model-Avg	0.28	0.24	0.06	0.29	**0.48**	0.20	0.09	0.14	0.06	0.14	0.06	0.06	0.02	0.03	0.15

Table 62: Comparison of unsupervised MIAs on the subsets of the SFT domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold.

Subset	Model	Likelihood-based										Pert.	Prompt		Out.-Dist.		Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	PAC	ReCaLL	EMMIA	CAMIA	Neigh.	Guided	DCQ	CDD	SelfCrit†	
CoT-GSM8K	1B	0.45	**0.55**	0.22	0.43	0.10	0.19	0.30	0.00	0.25	0.14	0.01	0.06	0.04	0.03	0.02	0.19
	7B	0.94	0.96	0.46	**0.97**	0.24	0.91	0.38	0.00	0.97	0.95	0.11	0.02	0.07	0.04	0.03	0.47
	13B	**1.00**	1.00	0.59	1.00	0.09	0.96	0.41	0.00	1.00	0.97	0.06	0.03	0.02	0.05	0.02	0.48
	32B	**1.00**	1.00	0.63	1.00	0.21	0.98	0.52	0.00	0.05	0.95	0.16	0.03	0.09	0.06	0.01	0.45
	Model-Avg	0.85	**0.88**	0.48	0.85	0.16	0.76	0.41	0.00	0.57	0.75	0.09	0.03	0.05	0.04	0.02	0.40
CoT-MATH	1B	0.14	0.15	0.03	**0.22**	0.05	0.13	0.08	0.05	0.05	0.14	0.06	0.05	0.04	0.06	0.02	0.08
	7B	**0.37**	0.34	0.20	0.34	0.08	0.29	0.11	0.03	0.05	0.11	0.11	0.06	0.04	0.02	0.06	0.15
	13B	**0.64**	0.62	0.24	0.57	0.04	0.45	0.20	0.01	0.05	0.27	0.15	0.01	0.08	0.06	0.08	0.23
	32B	0.42	0.37	0.23	**0.48**	0.09	0.43	0.16	0.06	0.03	0.39	0.18	0.03	0.03	0.05	0.04	0.20
	Model-Avg	0.39	0.37	0.17	**0.40**	0.06	0.33	0.14	0.04	0.04	0.23	0.12	0.04	0.05	0.05	0.05	0.17

Table 63: Comparison of unsupervised MIAs on the subsets of the RLVR domain of OLMO-DETECT, evaluated via TPR@5% FPR. Methods are grouped by family as in Table 50. **Model-Avg** averages across model sizes, while **Method-Avg** averages the methods with a value in that row. The best-performing method in each row is highlighted in bold. †SelfCrit is evaluated only on DPO and RLVR.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.59	0.64	0.65	0.67	0.64
Sup-CAMIA	In-domain	0.51	0.55	0.54	0.56	0.54
	LODO	**0.55**	**0.64**	**0.58**	**0.57**	**0.58**
MIA-Tuner	In-domain	0.45	0.45	0.53	0.55	0.50
	LODO	0.47	0.51	0.49	0.47	0.49

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.64	−0.02	−0.06	−0.06	−0.05	−0.05	−0.10	−0.04	−0.10	−0.18	−0.10
FSD (Zlib)	0.64	−0.01	−0.06	−0.04	−0.05	−0.04	−0.09	−0.05	−0.08	−0.19	−0.10
FSD (Lower)	0.63	−0.12	−0.09	−0.12	−0.09	−0.10	−0.09	−0.11	−0.11	−0.13	−0.11
FSD (MinK)	0.61	−0.07	−0.07	−0.07	−0.01	−0.05	−0.08	−0.04	−0.07	−0.18	−0.09
FSD (MinK++)	0.53	−0.03	−0.02	−0.01	−0.03	−0.02	−0.01	−0.03	−0.04	−0.15	−0.06

Table 64: Supervised MIAs on the DCLM-Baseline domain of OLMO-DETECT, evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.49	0.49	0.50	0.51	0.50
Sup-CAMIA	In-domain	**0.55**	0.47	0.52	0.50	0.51
	LODO	0.50	0.52	0.48	0.55	0.51
MIA-Tuner	In-domain	0.54	**0.57**	**0.56**	0.57	**0.56**
	LODO	0.55	0.51	0.39	**0.57**	0.51

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.48	**+0.07**	**+0.04**	**+0.05**	**+0.03**	**+0.05**	**+0.08**	**+0.05**	**+0.08**	**+0.07**	**+0.07**
FSD (Zlib)	0.50	**+0.05**	**+0.02**	**+0.01**	−0.01	**+0.02**	**+0.04**	**+0.03**	**+0.05**	**+0.05**	**+0.04**
FSD (Lower)	0.49	**+0.07**	**+0.04**	−0.00	**+0.02**	**+0.03**	**+0.04**	**+0.00**	−0.01	**+0.00**	**+0.01**
FSD (MinK)	0.50	**+0.06**	**+0.01**	−0.00	**+0.02**	**+0.02**	**+0.05**	**+0.01**	**+0.04**	**+0.04**	**+0.03**
FSD (MinK++)	0.50	**+0.04**	**+0.01**	−0.01	**+0.02**	**+0.01**	**+0.01**	**+0.01**	−0.01	**+0.06**	**+0.02**

Table 65: Supervised MIAs on the peS2o domain of OLMO-DETECT, evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.56	0.57	0.58	0.64	0.59
Sup-CAMIA	In-domain	**0.60**	**0.59**	**0.59**	**0.58**	**0.59**
	LODO	0.54	0.56	0.53	0.56	0.55
MIA-Tuner	In-domain	0.57	0.58	0.52	0.54	0.55
	LODO	0.59	0.54	0.56	0.51	0.55

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.53	**+0.04**	**+0.03**	**+0.07**	**+0.09**	**+0.06**	**+0.03**	**+0.04**	**+0.02**	**+0.02**	**+0.03**
FSD (Zlib)	0.59	**+0.02**	**+0.00**	**+0.02**	**+0.01**	**+0.01**	**+0.00**	−0.01	−0.05	−0.04	−0.02
FSD (Lower)	0.60	−0.04	−0.04	−0.01	−0.07	−0.04	−0.10	−0.04	−0.11	−0.07	−0.08
FSD (MinK)	0.56	**+0.01**	−0.00	**+0.02**	**+0.04**	**+0.02**	**+0.01**	**+0.01**	−0.02	−0.02	−0.01
FSD (MinK++)	0.52	−0.05	−0.01	**+0.02**	**+0.00**	−0.01	−0.01	−0.00	−0.04	−0.07	−0.03

Table 66: Supervised MIAs on the OpenWebMath domain of OLMO-DETECT, evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.65	0.65	0.64	0.63	0.64
Sup-CAMIA	In-domain	0.57	0.54	0.56	0.56	0.56
	LODO	**0.64**	**0.69**	**0.67**	**0.67**	**0.67**
MIA-Tuner	In-domain	0.61	0.54	0.53	0.48	0.54
	LODO	0.46	0.55	0.44	0.50	0.49

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.59	+**0.03**	+**0.01**	−0.00	−0.02	+**0.01**	+**0.02**	−0.01	−0.11	−0.01	−0.03
FSD (Zlib)	0.64	−0.02	−0.01	−0.01	−0.05	−0.02	−0.03	−0.06	−0.16	−0.03	−0.07
FSD (Lower)	0.60	+**0.02**	−0.00	−0.02	−0.02	−0.01	+**0.04**	−0.03	−0.03	−0.05	−0.02
FSD (MinK)	0.61	−0.03	−0.04	−0.06	−0.09	−0.05	−0.05	−0.01	−0.12	−0.04	−0.05
FSD (MinK++)	0.49	+**0.00**	−0.04	−0.00	−0.01	−0.01	+**0.03**	+**0.06**	−0.02	+**0.02**	+**0.02**

Table 67: Supervised MIAs on the StarCoder domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.76	0.89	0.99	0.99	0.91
Sup-CAMIA	In-domain	**0.82**	**0.94**	**1.00**	**1.00**	**0.94**
	LODO	0.68	0.85	0.77	0.90	0.80
MIA-Tuner	In-domain	0.51	0.51	0.50	0.51	0.51
	LODO	0.49	0.51	0.51	0.52	0.51

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.89	−0.05	−0.08	+**0.01**	−0.06	−0.04	−0.16	−0.10	−0.01	−0.14	−0.10
FSD (Zlib)	0.91	−0.07	−0.10	+**0.00**	−0.06	−0.06	−0.19	−0.12	−0.01	−0.13	−0.11
FSD (Lower)	0.85	−0.03	+**0.01**	+**0.01**	+**0.01**	−0.00	−0.13	−0.05	−0.04	−0.04	−0.06
FSD (MinK)	0.89	−0.12	−0.27	−0.06	−0.13	−0.15	−0.19	−0.08	−0.01	−0.12	−0.10
FSD (MinK++)	0.55	−0.02	−0.02	−0.02	−0.09	−0.04	−0.06	+**0.00**	−0.00	−0.05	−0.03

Table 68: Supervised MIAs on the GSM8K subset of OLMO-DETECT's Math Mix domain, evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.59	0.64	0.67	0.70	0.65
Sup-CAMIA	In-domain	0.61	0.61	**0.72**	**0.68**	0.66
	LODO	0.59	0.61	0.56	0.50	0.56
MIA-Tuner	In-domain	**0.69**	**0.72**	0.70	0.55	**0.66**
	LODO	0.51	0.46	0.44	0.54	0.49

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.59	+**0.01**	+**0.06**	+**0.11**	−0.00	+**0.05**	−0.08	+**0.03**	+**0.01**	−0.07	−0.03
FSD (Zlib)	0.65	−0.02	+**0.03**	+**0.03**	−0.08	−0.01	−0.11	−0.02	−0.08	−0.14	−0.09
FSD (Lower)	0.53	−0.01	−0.01	−0.03	−0.07	−0.03	+**0.01**	+**0.06**	+**0.08**	−0.08	+**0.02**
FSD (MinK)	0.60	+**0.01**	+**0.03**	+**0.04**	−0.01	+**0.02**	−0.08	+**0.01**	−0.01	−0.08	−0.04
FSD (MinK++)	0.57	+**0.00**	+**0.00**	−0.02	−0.03	−0.01	−0.03	+**0.01**	−0.02	−0.03	−0.02

Table 69: Supervised MIAs on the Stack Exchange domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.64	0.70	0.61	0.64	0.65
Sup-CAMIA	In-domain	0.51	**0.73**	**0.66**	**0.55**	**0.61**
	LODO	**0.55**	0.61	0.56	0.49	0.55
MIA-Tuner	In-domain	0.53	0.55	0.50	0.52	0.53
	LODO	0.50	0.46	0.48	0.51	0.49

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.66	+**0.04**	−0.05	+**0.02**	+**0.06**	+**0.02**	+**0.06**	−0.04	+**0.02**	+**0.04**	+**0.02**
FSD (Zlib)	0.65	+**0.01**	+**0.02**	+**0.08**	+**0.05**	+**0.04**	+**0.04**	+**0.02**	+**0.09**	+**0.04**	+**0.05**
FSD (Lower)	0.51	−0.03	+**0.01**	−0.05	−0.00	−0.02	−0.00	−0.04	−0.02	+**0.02**	−0.01
FSD (MinK)	0.66	−0.02	−0.07	+**0.02**	+**0.06**	−0.00	+**0.03**	−0.06	+**0.02**	+**0.05**	+**0.01**
FSD (MinK++)	0.72	−0.10	−0.06	−0.03	−0.01	−0.05	−0.06	−0.07	−0.02	−0.02	−0.04

Table 70: Supervised MIAs on the SFT domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.55	0.39	0.47	0.53	0.48
Sup-CAMIA	In-domain	**0.60**	**0.58**	**0.58**	**0.57**	**0.58**
	LODO	0.40	0.42	0.40	0.44	0.42
MIA-Tuner	In-domain	0.49	0.55	0.56	0.47	0.52
	LODO	0.41	0.44	0.42	0.48	0.44

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.42	−0.04	−0.01	−0.04	−0.02	−0.03	−0.03	−0.01	−0.06	−0.01	−0.03
FSD (Zlib)	0.48	−0.13	−0.02	−0.13	−0.04	−0.08	−0.12	−0.01	−0.15	−0.06	−0.09
FSD (Lower)	0.60	−0.08	**+0.02**	−0.03	−0.06	−0.04	−0.06	**+0.05**	−0.03	−0.10	−0.04
FSD (MinK)	0.43	−0.07	**+0.00**	−0.04	−0.02	−0.03	−0.06	−0.00	−0.06	−0.01	−0.03
FSD (MinK++)	0.44	−0.08	−0.01	−0.03	**+0.00**	−0.03	−0.07	−0.01	−0.02	**+0.02**	−0.02

Table 71: Supervised MIAs on the UF-Wildchat subset of OLMO-DETECT's DPO domain, evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.79	0.88	0.94	0.91	0.88
Sup-CAMIA	In-domain	**0.73**	**0.86**	**0.91**	0.90	**0.85**
	LODO	0.62	0.83	0.68	**0.94**	0.77
MIA-Tuner	In-domain	0.56	0.54	0.53	0.54	0.54
	LODO	0.51	0.48	0.48	0.52	0.50

Method	Base AUC	In-domain ΔAUC					LODO ΔAUC				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.89	−0.05	−0.21	−0.12	−0.31	−0.17	−0.02	−0.02	−0.03	−0.07	−0.03
FSD (Zlib)	0.88	−0.11	−0.25	−0.14	−0.31	−0.20	−0.05	−0.02	−0.02	−0.05	−0.04
FSD (Lower)	0.79	−0.04	−0.11	−0.05	−0.14	−0.08	−0.06	−0.07	−0.01	−0.01	−0.03
FSD (MinK)	0.90	−0.19	−0.59	−0.48	−0.59	−0.46	−0.10	−0.07	−0.04	−0.11	−0.08
FSD (MinK++)	0.66	−0.02	−0.07	−0.02	−0.14	−0.06	−0.01	−0.02	−0.00	−0.06	−0.02

Table 72: Supervised MIAs on the RLVR domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via AUC. **Top:** AUCs of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its AUC gain (ΔAUC) over the base MIA (in parentheses) on this domain. **Base AUC**: the base MIA's AUC on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B				7B				13B				32B				Model-Avg			
		Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All	Pre	Mid	Post	All
Zlib (unsup.)	—	0.09	0.17	0.18	0.15	0.13	0.33	0.32	0.26	0.13	0.57	0.34	0.35	0.16	0.58	0.31	0.35	0.13	0.41	0.29	0.28
Sup-CAMIA	In-domain	**0.08**	**0.24**	**0.19**	**0.17**	0.07	**0.38**	**0.30**	**0.25**	**0.08**	**0.57**	**0.36**	**0.34**	0.06	**0.56**	0.27	**0.30**	0.07	**0.44**	**0.28**	**0.26**
	LODO	0.07	0.09	0.08	0.08	0.06	0.22	0.25	0.18	0.07	0.07	0.11	0.08	**0.10**	0.24	**0.28**	0.21	**0.07**	0.16	0.18	0.14
MIA-Tuner	In-domain	0.07	0.13	0.10	0.10	**0.08**	0.20	0.09	0.12	0.04	0.15	0.08	0.09	0.06	0.07	0.06	0.06	0.06	0.14	0.08	0.09
	LODO	0.07	0.05	0.05	0.06	0.04	0.07	0.05	0.05	0.04	0.05	0.06	0.05	0.05	0.06	0.06	0.06	0.05	0.06	0.05	0.05

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.25	−0.00	−0.06	+**0.01**	−0.10	−0.04	−0.02	−0.05	−0.02	−0.13	−0.06
FSD (Zlib)	0.28	−0.03	−0.10	−0.06	−0.12	−0.08	−0.04	−0.07	−0.04	−0.14	−0.07
FSD (Lower)	0.18	−0.02	−0.02	−0.00	−0.05	−0.02	−0.04	−0.02	−0.03	−0.05	−0.04
FSD (MinK)	0.26	−0.05	−0.15	−0.11	−0.15	−0.11	−0.05	−0.08	−0.03	−0.14	−0.07
FSD (MinK++)	0.12	−0.04	−0.04	−0.01	−0.06	−0.04	−0.06	−0.04	−0.01	−0.04	−0.04

Table 73: Supervised MIAs on OLMO-DETECT, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on the evaluated domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Pre/Mid/Post** macro-average over each stage's domains and **All** macro-averages the three stages, as in Table 4. **Bottom:** FSD's gain (Δ) over its base MIA (in parentheses) under the **All** aggregation; **Base TPR@5% FPR** is the base MIA's own score averaged across model sizes, and positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.06	0.11	0.13	0.18	0.12
Sup-CAMIA	In-domain	0.01	0.05	**0.04**	0.06	0.04
	LODO	**0.05**	**0.08**	0.04	**0.17**	**0.09**
MIA-Tuner	In-domain	0.01	0.01	0.00	0.06	0.02
	LODO	0.02	0.02	0.03	0.01	0.02

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.15	−0.05	−0.10	−0.08	−0.13	−0.09	−0.06	−0.01	−0.09	−0.17	−0.08
FSD (Zlib)	0.12	−0.01	−0.07	−0.10	−0.07	−0.06	−0.02	+**0.03**	−0.04	−0.11	−0.04
FSD (Lower)	0.14	−0.07	−0.04	−0.02	−0.03	−0.04	−0.07	−0.06	+**0.00**	−0.04	−0.04
FSD (MinK)	0.19	−0.11	−0.19	−0.15	−0.11	−0.14	−0.09	−0.11	−0.14	−0.18	−0.13
FSD (MinK++)	0.06	+**0.01**	−0.01	+**0.01**	−0.06	−0.01	+**0.04**	−0.04	−0.01	−0.09	−0.03

Table 74: Supervised MIAs on the DCLM-Baseline domain of OLMO-DETECT, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.07	0.07	0.07	0.09	0.07
Sup-CAMIA	In-domain	0.08	0.04	0.05	0.08	0.06
	LODO	0.05	0.03	**0.06**	0.05	0.05
MIA-Tuner	In-domain	**0.09**	0.08	0.06	0.09	0.08
	LODO	0.07	**0.12**	0.01	**0.13**	**0.08**

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.07	−0.02	−0.02	−0.03	−0.04	−0.03	−0.01	−0.01	+**0.05**	−0.01	+**0.01**
FSD (Zlib)	0.07	−0.02	+0.00	−0.01	−0.05	−0.02	−0.02	−0.02	+**0.03**	−0.04	−0.01
FSD (Lower)	0.07	+**0.00**	−0.03	−0.03	−0.01	−0.02	+**0.01**	−0.03	−0.03	−0.04	−0.02
FSD (MinK)	0.08	−0.03	−0.03	−0.04	−0.05	−0.04	−0.02	−0.03	+**0.00**	−0.03	−0.02
FSD (MinK++)	0.07	−0.04	−0.00	−0.03	−0.01	−0.02	−0.02	−0.00	+**0.01**	+**0.00**	−0.00

Table 75: Supervised MIAs on the peS2o domain of OLMO-DETECT, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.07	0.07	0.10	0.17	0.10
Sup-CAMIA	In-domain	0.10	0.06	**0.10**	0.04	**0.07**
	LODO	0.05	0.09	0.03	**0.09**	0.06
MIA-Tuner	In-domain	0.05	**0.09**	0.06	0.05	0.06
	LODO	**0.11**	0.01	0.06	0.07	0.06

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.06	+**0.04**	+**0.06**	+**0.04**	+**0.06**	+**0.05**	+**0.01**	+0.00	−0.02	−0.05	−0.02
FSD (Zlib)	0.10	+**0.02**	+**0.04**	+**0.02**	−0.01	+**0.02**	−0.01	−0.04	−0.05	−0.07	−0.04
FSD (Lower)	0.06	−0.00	−0.05	+**0.01**	−0.01	−0.01	−0.03	−0.03	−0.02	−0.00	−0.02
FSD (MinK)	0.06	+**0.04**	+**0.05**	+**0.03**	+**0.11**	+**0.05**	+**0.01**	−0.00	−0.02	−0.03	−0.01
FSD (MinK++)	0.12	−0.03	−0.10	−0.07	−0.09	−0.07	+**0.01**	−0.09	−0.08	−0.06	−0.06

Table 76: Supervised MIAs on the OpenWebMath domain of OLMO-DETECT, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.17	0.27	0.24	0.19	0.22
Sup-CAMIA	In-domain	0.12	**0.13**	0.13	0.08	**0.12**
	LODO	0.12	0.04	**0.14**	**0.10**	0.10
MIA-Tuner	In-domain	**0.12**	0.11	0.02	0.05	0.08
	LODO	0.09	0.02	0.05	0.01	0.04

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.08	−0.03	−0.01	−0.01	−0.02	−0.02	+**0.06**	−0.01	−0.03	−0.02	+**0.00**
FSD (Zlib)	0.22	−0.05	−0.17	−0.11	−0.05	−0.09	−0.03	−0.18	−0.15	−0.19	−0.14
FSD (Lower)	0.07	+**0.06**	+**0.02**	−0.05	+**0.06**	+**0.02**	+**0.03**	+**0.03**	−0.05	−0.02	+**0.00**
FSD (MinK)	0.08	−0.06	−0.05	−0.03	−0.03	−0.04	+**0.07**	+**0.02**	−0.03	−0.02	+**0.01**
FSD (MinK++)	0.08	−0.11	−0.02	+**0.00**	−0.02	−0.04	−0.04	+**0.03**	−0.01	+**0.02**	−0.00

Table 77: Supervised MIAs on the StarCoder domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.25	0.54	0.97	0.96	0.68
Sup-CAMIA	In-domain	**0.39**	**0.69**	**0.97**	**0.97**	**0.76**
	LODO	0.13	0.38	0.07	0.43	0.25
MIA-Tuner	In-domain	0.06	0.05	0.06	0.06	0.06
	LODO	0.05	0.07	0.05	0.06	0.06

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.61	−0.02	−0.06	+**0.05**	−0.16	−0.05	−0.17	−0.22	−0.05	−0.53	−0.24
FSD (Zlib)	0.68	−0.08	−0.15	+**0.01**	−0.21	−0.11	−0.20	−0.29	−0.03	−0.49	−0.25
FSD (Lower)	0.38	−0.01	+**0.07**	+**0.16**	+**0.05**	+**0.06**	−0.12	+**0.01**	−0.09	−0.18	−0.09
FSD (MinK)	0.63	−0.11	−0.39	−0.21	−0.32	−0.26	−0.20	−0.25	−0.09	−0.41	−0.24
FSD (MinK++)	0.16	−0.04	−0.03	−0.04	−0.09	−0.05	−0.10	−0.07	−0.03	−0.07	−0.07

Table 78: Supervised MIAs on the GSM8K subset of OLMO-DETECT's Math Mix domain, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.08	0.12	0.17	0.20	0.14
Sup-CAMIA	In-domain	0.08	0.06	0.17	**0.16**	0.12
	LODO	0.05	0.06	0.08	0.06	0.06
MIA-Tuner	In-domain	**0.19**	**0.36**	**0.24**	0.09	**0.22**
	LODO	0.06	0.08	0.04	0.05	0.06

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.07	+**0.02**	+**0.08**	+**0.15**	+**0.02**	+**0.07**	−0.03	−0.01	+**0.02**	−0.04	−0.02
FSD (Zlib)	0.14	−0.00	+**0.02**	−0.03	−0.08	−0.02	−0.05	−0.08	−0.04	−0.15	−0.08
FSD (Lower)	0.13	−0.07	−0.03	−0.00	−0.12	−0.06	−0.06	−0.00	−0.06	−0.11	−0.06
FSD (MinK)	0.08	+**0.00**	−0.01	+**0.08**	−0.02	+**0.01**	−0.02	−0.01	+**0.02**	−0.07	−0.02
FSD (MinK++)	0.07	−0.02	−0.00	−0.00	−0.04	−0.01	−0.02	+**0.03**	+**0.04**	−0.04	+**0.00**

Table 79: Supervised MIAs on the Stack Exchange domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.14	0.29	0.14	0.15	0.18
Sup-CAMIA	In-domain	**0.09**	0.20	**0.14**	0.08	**0.13**
	LODO	0.06	**0.23**	0.05	**0.10**	0.11
MIA-Tuner	In-domain	0.09	0.10	0.07	0.04	0.07
	LODO	0.05	0.03	0.07	0.03	0.04

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.27	+**0.08**	+**0.03**	+**0.16**	+**0.01**	+**0.07**	+**0.19**	−0.01	+**0.03**	−0.08	+**0.03**
FSD (Zlib)	0.18	+**0.12**	−0.00	+**0.07**	+**0.12**	+**0.08**	+**0.18**	+**0.08**	+**0.07**	+**0.05**	+**0.09**
FSD (Lower)	0.09	−0.03	−0.01	+**0.00**	−0.02	−0.02	−0.05	−0.05	+**0.01**	−0.03	−0.03
FSD (MinK)	0.27	+**0.09**	+**0.01**	+**0.13**	−0.09	+**0.04**	+**0.14**	−0.06	+**0.08**	−0.12	+**0.01**
FSD (MinK++)	0.36	−0.12	−0.09	+**0.01**	−0.20	−0.10	−0.28	−0.21	−0.04	−0.15	−0.17

Table 80: Supervised MIAs on the SFT domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.06	0.03	0.06	0.09	0.06
Sup-CAMIA	In-domain	**0.21**	0.06	**0.21**	0.05	**0.13**
	LODO	0.06	0.06	0.11	**0.08**	0.08
MIA-Tuner	In-domain	0.05	**0.08**	0.12	0.06	0.08
	LODO	0.06	0.07	0.07	0.07	0.07

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.04	−0.00	−0.01	−0.02	+**0.01**	−0.01	−0.02	+**0.01**	−0.01	+**0.01**	−0.00
FSD (Zlib)	0.06	+**0.03**	−0.03	−0.04	−0.01	−0.01	+**0.00**	−0.01	−0.03	−0.03	−0.02
FSD (Lower)	0.15	−0.01	+**0.01**	−0.01	−0.02	−0.01	+**0.01**	+**0.05**	+**0.00**	−0.03	+**0.01**
FSD (MinK)	0.05	−0.02	−0.01	−0.02	+**0.05**	+**0.00**	−0.01	+**0.03**	−0.02	+**0.04**	+**0.01**
FSD (MinK++)	0.04	+**0.01**	−0.01	+**0.01**	+**0.03**	+**0.01**	−0.01	+**0.01**	−0.01	+**0.05**	+**0.01**

Table 81: Supervised MIAs on the UF-Wildchat subset of OLMO-DETECT's DPO domain, evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

Method	Setting	1B	7B	13B	32B	Model-Avg
Zlib (unsup.)	—	0.35	0.65	0.81	0.68	0.62
Sup-CAMIA	In-domain	**0.27**	**0.64**	**0.72**	**0.67**	**0.58**
	LODO	0.12	0.45	0.18	0.66	0.35
MIA-Tuner	In-domain	0.15	0.10	0.07	0.07	0.10
	LODO	0.04	0.04	0.03	0.08	0.05

Method	Base TPR@5% FPR	In-domain Δ					LODO Δ				
		1B	7B	13B	32B	Model-Avg	1B	7B	13B	32B	Model-Avg
FSD (PPL)	0.62	−0.06	−0.51	−0.33	−0.60	−0.37	−0.07	−0.05	−0.11	−0.06	−0.07
FSD (Zlib)	0.62	−0.23	−0.53	−0.43	−0.59	−0.44	−0.08	−0.01	−0.09	−0.03	−0.05
FSD (Lower)	0.32	−0.03	−0.18	−0.18	−0.31	−0.17	−0.06	−0.09	−0.01	**+0.09**	−0.02
FSD (MinK)	0.63	−0.21	−0.57	−0.77	−0.70	−0.56	−0.22	−0.19	−0.13	−0.23	−0.19
FSD (MinK++)	0.11	−0.01	−0.07	−0.01	−0.06	−0.04	−0.01	−0.01	+0.00	−0.02	−0.01

Table 82: Supervised MIAs on the RLVR domain of OLMO-DETECT (macro-averaged over its subsets), evaluated via TPR@5% FPR. **Top:** scores of Sup-CAMIA and MIA-Tuner under the **In-domain** setting (trained on this domain's dev set) and the **LODO** setting (trained on the other eight domains' dev sets); Zlib, the strongest unsupervised MIA, is repeated as a reference and is excluded from the bold, which marks the best supervised entry per column. **Bottom:** FSD measures the change in a base MIA's score after fine-tuning the target model on non-members, so we report its gain (Δ) over the base MIA (in parentheses) on this domain. **Base TPR@5% FPR**: the base MIA's own score on this domain averaged across model sizes; positive gains are bolded. **Model-Avg** averages across model sizes.

M Effect of the Labeled-Data Budget on Supervised MIAs

The supervised methods in Section 5.2 are trained on each domain's dev split, i.e., 10% of its members and non-members. To test whether this budget limits their performance, we measure Sup-CAMIA as a function of the number of labeled examples. For each domain and model size, we hold out half of the members and half of the non-members as a fixed evaluation set and train on nested subsets of the remaining half containing 10%, 20%, 30%, 40%, and 50% of the domain's members and non-members; the 10% point matches the budget of the main tables. We repeat this over five random splits and report the mean AUC on the held-out half. Because the evaluation set comprises half of each domain rather than the 90% test split, absolute AUCs are not comparable to those in the main tables; only the trend is relevant.

Table 83 reports the results. Raising the budget from 10% to 50% increases the macro-averaged AUC by 0.04–0.05 at every model size. The gains are confined to the smallest pre-training domains (DCLM-Baseline, OpenWebMath, and StarCoder, with at most 151 members), where the 10% budget provides only 15–16 labeled members and non-members; peS2o and DPO stay within 0.03 of their 10% value, and GSM8K and CoT-GSM8K are saturated at every budget. A five-fold larger labeled set therefore does not change the conclusion of Section 5.2: supervision from the matched distribution adds little beyond unsupervised membership signals.

N Domain-Unknown Evaluation

The main text evaluates MIAs *given the domain*: each domain is scored separately and the AUCs are macro-averaged, which assumes the auditor knows which domain a text comes from. In practice that is rarely true. To simulate the realistic case, the *domain-unknown* setting pools members and non-members from all nine domains and computes a single AUC, so a score must separate members from non-members in general rather than within a known domain.[2] Table 84 contrasts the two settings on the same subsampled data, 100 members and 100 non-members per domain over 20 draws, with supervised MIAs retrained on the pooled dev sets. For comparison on a common metric we also report accuracy, with a threshold tuned within each evaluated domain in the first setting and a single threshold tuned on the pooled data in the second.

Pooling lowers AUC for 12 of the 15 unsupervised MIAs. The three that rise, ReCaLL, Guided and DCQ, are at or below chance in both settings, so the increase carries no information. The largest drops fall on Zlib (0.64 to 0.59), PPL (0.60 to 0.56) and EMMIA (0.55 to 0.51), while CAMIA is unaffected (0.62 in both), making it the strongest method once the domain is unknown. Averaged over unsupervised MIAs the cost is 0.02 AUC. Supervised MIAs pay more, but unevenly: Sup-CAMIA falls from 0.66 to 0.60, having lost the ability to calibrate to a single domain, whereas MIA-Tuner is almost unchanged at 0.56 to 0.55. Accuracy tells the same story, with the unsupervised average flat at 0.55 and Sup-CAMIA falling from 0.62 to 0.58. Crucially the ranking survives: Spearman $\rho = 0.95$ between the two settings across the 17 MIAs. Absolute performance therefore declines when the domain is unknown, but the conclusions of the main text do not depend on knowing it.

O Per-Domain Matched-vs-Shifted Results

Table 85 reports the per-domain changes behind the aggregate counts of Table 7. Each entry is the AUC on OLMo-Detect (Shifted) minus the AUC on OLMo-Detect, averaged over the four model sizes, with significance from a paired instance-level bootstrap. Two patterns are visible directly. The direction of the change is a property of the method–domain pair rather than of the domain: OpenWebMath inflates PPL, MinK, CAMIA and EMMIA but degrades Zlib, Lower and PAC. And the largest change of every MIA falls in a pre-training domain, most often OpenWebMath (7 of 14 MIAs), then StarCoder (4) and peS2o (3), while DCLM never carries a method's largest change despite being shifted on the same axes.

[2] This differs from the LODO setting: the test instance still comes from a domain the MIA was tuned on, we simply do not know which one.

Domain	N	1B					7B					13B					32B				
		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
DCLM-Baseline	150	0.52	0.57	0.56	0.57	0.58	0.56	0.58	0.62	0.62	0.63	0.58	0.56	0.59	0.60	0.60	0.58	0.60	0.63	0.63	0.64
peS2o	268	0.52	0.53	0.55	0.55	0.55	0.52	0.55	0.55	0.55	0.55	0.49	0.50	0.49	0.49	0.48	0.49	0.49	0.50	0.49	0.49
OpenWebMath	151	0.54	0.57	0.60	0.62	0.62	0.52	0.55	0.58	0.61	0.61	0.52	0.57	0.58	0.60	0.60	0.55	0.60	0.61	0.63	0.63
StarCoder	156	0.61	0.63	0.65	0.65	0.66	0.59	0.61	0.65	0.67	0.67	0.59	0.61	0.64	0.65	0.66	0.61	0.64	0.66	0.66	0.68
Math Mix	1040	0.83	0.83	0.84	0.84	0.84	0.94	0.94	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Stack Exchange	430	0.58	0.58	0.58	0.58	0.59	0.57	0.59	0.60	0.61	0.62	0.63	0.64	0.67	0.69	0.70	0.60	0.63	0.65	0.67	0.68
SFT	164–225	0.59	0.62	0.62	0.64	0.64	0.67	0.76	0.77	0.78	0.77	0.58	0.64	0.64	0.64	0.65	0.58	0.59	0.62	0.63	0.64
DPO	110–113	0.61	0.61	0.63	0.64	0.65	0.54	0.59	0.60	0.61	0.61	0.64	0.65	0.66	0.67	0.67	0.58	0.59	0.61	0.61	0.63
RLVR	290	0.73	0.78	0.80	0.80	0.80	0.86	0.87	0.87	0.88	0.89	0.94	0.95	0.95	0.95	0.95	0.90	0.92	0.92	0.92	0.92
Macro-Avg		**0.61**	**0.63**	**0.65**	**0.65**	**0.66**	**0.64**	**0.67**	**0.69**	**0.70**	**0.70**	**0.66**	**0.68**	**0.69**	**0.70**	**0.70**	**0.65**	**0.67**	**0.69**	**0.69**	**0.70**

Table 83: AUC of Sup-CAMIA as a function of the labeled-data budget on OLMO-DETECT. For each subset and model size, half of the members and half of the non-members are held out as a fixed evaluation set; Sup-CAMIA is trained on nested subsets of the remaining half containing 10%–50% of the subset's members and non-members (10% equals the dev-set budget of the main tables). Entries are the mean AUC on the held-out half over five random splits, averaged over the domain's subsets (DPO over the chosen and rejected responses); N is the number of members (equal to the number of non-members) in the domain. SFT and DPO members come from the size-specific post-training mixes (SFT: one mix shared by 1B and 32B and another by 7B and 13B; DPO: one preference mix per size), so their N varies across model sizes and is given as a range. Macro-Avg is the mean over the nine domains. At the subset level the standard deviation across splits is below 0.05 in 211 of 260 cases; it reaches 0.17 for the smallest subsets (at most 100 members) at the 10%–20% budgets and is at most 0.09 at budgets of 30% or more. Absolute values are not comparable to Table 5, which uses the 90% test split.

P NON-MEMBER VERIFICATION FOR SMOLLM2

SmolLM2-1.7B is trained on DCLM together with FineWeb-Edu, both publicly released. Its members therefore overlap OLMO-DETECT's DCLM domain by construction, but its non-members are valid only if they are absent from *every* corpus SmolLM2 saw, not from DCLM alone. Because OLMO-DETECT's DCLM non-members are drawn from 25 Common Crawl dumps spanning 2013–2021, and FineWeb-Edu deduplicates within rather than across dumps, absence from DCLM does not imply absence from FineWeb-Edu.

We therefore applied the same 13-gram filter used in benchmark construction, with the same 0.20 threshold, to the whole of FineWeb-Edu: 2,410 files and 1.53×10^9 documents, sharded nine ways across the fleet. Eight of the 134 DCLM test non-members exceed the threshold, at overlaps of 0.855, 0.797, 0.790, 0.746, 0.746, 0.733, 0.661 and 0.616, together with one of the 16 development non-members. The remaining 126 reach at most 0.128, so the separation is unambiguous rather than borderline. The excluded indices are released with the benchmark.

We evaluate SmolLM2 on the 126 verified non-members, and re-evaluate OLMo 2-1B on exactly the same set so that the comparison in Table 8 is paired. Members are not filtered: a SmolLM2 member appearing in FineWeb-Edu is expected and harmless, since it remains a member of that model's training data.

Method	Given the domain		Domain unknown	
	AUC (macro)	Acc.	AUC (pooled)	Acc.
PPL^{L}	0.60	0.57	0.56	0.56
Zlib^{L}	**0.64**	**0.59**	0.59	0.57
Lower^{L}	0.61	0.57	0.59	0.57
MinK^{L}	0.61	0.58	0.58	0.57
MinK++^{L}	0.54	0.52	0.52	0.53
DCPDD^{L}	0.63	0.59	0.60	0.58
PAC^{L}	0.63	0.58	0.59	0.58
ReCaLL^{L}	0.48	0.52	0.50	0.53
EMMIA^{L}	0.55	—	0.51	—
CAMIA^{L}	0.62	0.59	**0.62**	**0.60**
Neigh.^{P}	0.57	0.53	0.55	0.54
Guided^{Pr}	0.49	0.48	0.49	0.50
DCQ^{Pr}	0.49	0.50	0.50	0.53
CDD^{O}	0.52	0.50	0.51	0.50
SelfCrit^{O†}	0.48	—	0.48	—
Method-Avg (unsup.)	0.56	0.55	0.55	0.55
Sup-CAMIA	**0.66**	**0.62**	**0.60**	**0.58**
MIA-Tuner	0.56	0.52	0.55	0.54

Table 84: Two evaluation settings on OLMO-DETECT, averaged over model sizes (1B–32B). Every domain is subsampled to 100 members and 100 non-members (20 random draws; multi-subset domains pooled over their subsets; methods, dev sets and hyperparameters exactly as in the main tables). **Given the domain**: AUC per domain, macro-averaged (the paper's protocol), and accuracy with a decision threshold tuned within the evaluated domain (2-fold). **Domain unknown**: one AUC over the pooled instances of all nine domains, and accuracy with a single threshold tuned on the pooled data; supervised MIAs are retrained on the pooled dev sets of all nine domains. Thresholds maximise accuracy on their tuning data. Best per column and block in bold. †SelfCrit is evaluated on DPO and RLVR only. Accuracies are omitted for EMMIA, which reports the best of its four base scores per subset and thus has no single score to threshold and for SelfCrit.

Method	DCLM	peS2o	OWM	SC	SE	SFT	DPO	Largest
PPL	$+\mathbf{0.06}^{\ddagger}$	$+\mathbf{0.22}^{\ddagger}$	$+\mathbf{0.29}^{\ddagger}$	$+\mathbf{0.04}^{\dagger}$	$-\mathbf{0.06}^{\ddagger}$	$-\mathbf{0.08}^{\ddagger}$	-0.01	OWM
Zlib	$+0.01$	$-\mathbf{0.20}^{\ddagger}$	$-\mathbf{0.31}^{\ddagger}$	$+\mathbf{0.25}^{\ddagger}$	$+\mathbf{0.05}^{\ddagger}$	-0.01	$+0.01$	OWM
Lower	$+\mathbf{0.05}^{\ddagger}$	$+\mathbf{0.06}^{\ddagger}$	$-\mathbf{0.30}^{\ddagger}$	$+\mathbf{0.14}^{\ddagger}$	$+\mathbf{0.04}^{\ddagger}$	$+\mathbf{0.06}^{\ddagger}$	$+0.01$	OWM
MinK	$+0.02$	$+\mathbf{0.21}^{\ddagger}$	$+\mathbf{0.23}^{\ddagger}$	$+0.03$	-0.02	$-\mathbf{0.06}^{\ddagger}$	$+0.01$	OWM
MinK++	$-\mathbf{0.06}^{\ddagger}$	$-\mathbf{0.03}^{\dagger}$	-0.02	$-\mathbf{0.16}^{\ddagger}$	$+\mathbf{0.07}^{\ddagger}$	-0.02	$+0.01$	SC
DCPDD	$+0.03$	$+\mathbf{0.04}^{\ddagger}$	$-\mathbf{0.05}^{\ddagger}$	$+\mathbf{0.16}^{\ddagger}$	$+0.02$	$-\mathbf{0.10}^{\ddagger}$	$-\mathbf{0.04}^{\ddagger}$	SC
PAC	$+\mathbf{0.05}^{\dagger}$	$+\mathbf{0.09}^{\ddagger}$	$-\mathbf{0.20}^{\ddagger}$	$+\mathbf{0.19}^{\ddagger}$	-0.00	$+\mathbf{0.04}^{\dagger}$	$+\mathbf{0.09}^{\ddagger}$	OWM
ReCaLL	$+\mathbf{0.06}^{\ddagger}$	$+\mathbf{0.24}^{\ddagger}$	$+\mathbf{0.04}^{\ddagger}$	$+\mathbf{0.13}^{\ddagger}$	$+0.01$	$+0.02$	$+0.01$	peS2o
EMMIA	-0.01	$-\mathbf{0.17}^{\ddagger}$	$+\mathbf{0.42}^{\ddagger}$	$+\mathbf{0.06}^{\dagger}$	$-\mathbf{0.06}^{\ddagger}$	$+\mathbf{0.06}^{\ddagger}$	$+0.02$	OWM
CAMIA	$+0.04$	$+\mathbf{0.28}^{\ddagger}$	$+\mathbf{0.18}^{\ddagger}$	$+\mathbf{0.23}^{\ddagger}$	$-\mathbf{0.12}^{\ddagger}$	$+0.00$	$+\mathbf{0.03}^{\dagger}$	peS2o
Neigh.	$+0.03$	$+\mathbf{0.10}^{\ddagger}$	$+\mathbf{0.04}^{\dagger}$	$+\mathbf{0.12}^{\ddagger}$	$-\mathbf{0.03}^{\dagger}$	$+\mathbf{0.06}^{\ddagger}$	$+\mathbf{0.06}^{\ddagger}$	SC
Guided	$+0.03$	$-\mathbf{0.05}^{\ddagger}$	$+0.03$	$+\mathbf{0.10}^{\ddagger}$	$+0.00$	$-\mathbf{0.03}^{\dagger}$	-0.01	SC
DCQ	-0.02	$+\mathbf{0.09}^{\ddagger}$	$-\mathbf{0.03}^{\dagger}$	$+0.01$	$+\mathbf{0.03}^{\dagger}$	$+0.02$	-0.01	peS2o
CDD	$+0.01$	$+\mathbf{0.05}^{\ddagger}$	$+\mathbf{0.12}^{\ddagger}$	$+\mathbf{0.04}^{\ddagger}$	-0.01	-0.01	-0.01	OWM

Table 85: Per-domain effect of distribution shift on unsupervised MIAs, Δ = AUC on OLMo-Detect (Shifted) minus AUC on OLMo-Detect, averaged over the four model sizes. Significance is assessed with a paired instance-level bootstrap: † the 95% interval excludes zero, ‡ the 99% interval excludes zero; significant entries are in bold. **Largest**: the domain carrying that MIA's largest change. GSM8K and RLVR have no shifted variant. OWM: OpenWebMath; SC: StarCoder; SE: Stack Exchange. Aggregate counts are in Table 7.