OLMo-Detect: A Multi-Stage Benchmark for Verbatim Contamination Detection in Large Language Models

Anonymous ACL submission

Abstract

Verbatim contamination detection in large language models (LLMs) aims to determine whether a given instance appears verbatim 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 contaminated and uncontaminated splits, and lack of rigorous verification of uncontaminated splits. To address these limitations, we propose OLMo-Detect, a multi-stage benchmark constructed from the fully open OLMo 2 training pipeline. OLMo-Detect spans pre-training, mid-training, and post-training, explicitly aligns contaminated and uncontaminated splits along multiple axes, and rigorously verifies uncontaminated splits to minimize latent contamination. To assess method robustness under distributional shift, we further introduce OLMo-Detect (Shifted), a benchmark variant with distributional mismatches. We evaluate 15 state-of-the-art detection methods across the OLMo 2 model family, finding that: (i) no method exceeds an average AUC of 0.75 on OLMo-Detect; (ii) detection performance is strongest at mid-training and weakest at post-training, primarily reflecting domain composition rather than an inherent stage effect; (iii) detection performance plateaus at 13B and does not improve at 32B; and (iv) existing methods show varying behavior under distributional shift, with some inflating to near-perfect AUC. We release our benchmark and code at `https://anonymous.4open.science/r/OLMo-Detect-3E76`.

1 Introduction

Verbatim contamination detection in large language models (LLMs) (Brown et al., 2020; OpenAI et al., 2024b; Anthropic, 2024) aims to determine whether a given text sample appears *verbatim* 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; Fu et al., 2025; Tao et al., 2025; Chang et al., 2025; Kim et al., 2026), a fundamental limitation persists: **the absence of rigorous multi-stage benchmarks**.

As illustrated in Table 1, existing benchmarks suffer from three limitations: (i) **Limited coverage of training stages.** Modern LLM development (OLMo et al., 2025; Grattafiori et al., 2024; DeepSeek-AI 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 addressing post-training (Dong et al., 2024; Tao et al., 2025). To date, none covers mid-training or spans all three stages. (ii) **Insufficient distributional alignment between contaminated and uncontaminated splits.** Most existing benchmarks are constructed either in ways that introduce distributional shift (Shi et al., 2023; Ye et al., 2024; Zhang et al., 2024b; Dong et al., 2024; Fu et al., 2025) or without explicit control over distributional alignment (Tao et al., 2025; Kim et al., 2026). The only exception is OLMoMIA (Kim et al., 2026), which performs explicit alignment along semantic similarity. However, semantic similarity may not be the ideal axis for verbatim contamination detection, as the task fundamentally concerns surface-form overlap rather than semantic equivalence. (iii) **Lack of rigorous verification of uncontaminated splits.** Most existing benchmarks construct uncontaminated splits without verifying whether they share substantial n-gram overlap with contaminated splits, leading to potential latent contamination. While MIMIR (Duan et al., 2024) attempts explicit filtering via a 13-gram Bloom filter, uncontaminated splits in several domains still exhibit an over 30% 7-gram overlap with contaminated splits.

Benchmark	Stage(s)	Domain(s)	Contaminated Splits	Uncontaminated Splits	Assessments		
					C	A	V
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 (SFT)	HumanEval, GSM8K	Outputs of post-SFT LLMs	Outputs of base LLMs	✗	✗	✗
MIMIR (Duan et al., 2024)	Pre-training	The Pile: 7 domains, Dolma: 5 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 (RL)	AIME24, AIME25, GSM8K, K&K, SAT	Subset injection into a larger RL corpus	Non-injected data	✗	✗	✗
OLMoMIA (Kim et al., 2026)	Pre-training	Dolma	Train splits	Held-out data	✗	✓‡	✗
OLMo-Detect (ours)	Pre-training	4 domains	OLMo 2 training data	Distribution-matched uncontaminated data	✓	✓	✓
	Mid-training	2 domains					
	Post-training	3 domains					

Table 1: Comparison of existing benchmarks for verbatim contamination detection in LLMs. We assess each benchmark on three criteria: (i) **Coverage (C)** of all training stages (pre-training, mid-training, and post-training); (ii) **Explicit Alignment (A)** of distributions between contaminated and uncontaminated splits; (iii) **Rigorous Verification (V)** of uncontaminated splits. ✓† indicates that despite MIMIR's explicit verification, uncontaminated splits in several domains still retain over 30% 7-gram overlap with contaminated splits. ✓‡ denotes that OLMoMIA aligns along semantic similarity, which may be suboptimal, as the task fundamentally targets surface-form overlap.

Inspired by the fully open training pipeline of the OLMo 2 suite (OLMo et al., 2025), we propose OLMo-Detect, a multi-stage benchmark for verbatim contamination detection in LLMs. OLMo-Detect addresses all aforementioned limitations: (i) the OLMo 2 training corpus spans pre-training, mid-training, and post-training, enabling systematic analysis of how performance varies across training stages; (ii) contaminated and uncontaminated splits are explicitly aligned along three key axes to maximize distributional match: text quality, temporal range, and lexical similarity; and (iii) powered by infini-gram (Liu et al., 2024), we perform rigorous filtering of uncontaminated splits, achieving an average 7-gram overlap of 1.6% between contaminated and uncontaminated splits. Moreover, we introduce OLMo-Detect (Shifted), a benchmark variant with distributional mismatches. OLMo-Detect and OLMo-Detect (Shifted) serve complementary purposes: OLMo-Detect isolates genuine contamination signals under distributionally matched conditions, evaluating methods' intrinsic detection capability, while OLMo-Detect (Shifted) assesses robustness under distributional shift.

We evaluate 15 state-of-the-art detection methods across the OLMo 2 model family (1B, 7B, 13B, 32B), finding that: (i) no method exceeds an average AUC of 0.75 on OLMo-Detect, indicating that once distributional shift is minimized, true contamination signals become difficult to detect; (ii) detection performance is strongest at mid-training, moderate at pre-training, and weakest at post-training, primarily reflecting each stage's composition of domains with domain-specific structural framing (e.g., the Question/Reasoning/Answer template in GSM8K) rather than an inherent stage effect; (iii) detection performance improves up to 13B but plateaus at 32B, suggesting that detectability does not necessarily scale with model size as verbatim memorization capability does (Carlini et al., 2022); (iv) existing methods on OLMo-Detect (Shifted) show varying behavior, with several inflating to AUC > 0.95 (e.g., DCPDD) and some exhibiting degradation (e.g., Zlib).

2 Related Work

As summarized in Table 1, existing benchmarks can be grouped into three categories by construction strategy: temporal-cutoff (Shi et al., 2023; Ye et al., 2024; Zhang et al., 2024b; Fu et al., 2025), partition-based (Duan et al., 2024; Tao et al., 2025; Kim et al., 2026), and model-output-based (Dong et al., 2024). None satisfies all three criteria of coverage, explicit alignment, and rigorous verification. Details are provided in Appendix A.

Existing methods can be grouped into five categories by the signal they exploit: likelihood-based (Yeom et al., 2018; Carlini et al., 2021; Mattern et al., 2023; 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), ordering-based (Oren et al., 2023; Samuel et al., 2025), prompt-based (Golchin and Surdeanu, 2023; Chang et al., 2023; Deng et al., 2024), perturbation-based (Golchin and Surdeanu, 2025), and output-distribution-based (Dong et al., 2024; Tao et al., 2025). Details are provided in Appendix B.

3 Benchmark Construction

3.1 Motivation for Using OLMo 2

Constructing evaluation benchmarks requires fully open-source LLMs with publicly released training data, as contaminated splits require known training-data membership, and verifying uncontaminated splits requires substring lookup against the training corpus. Among existing fully 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 data, 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 the OLMo 2 suite, as it was the most recent LLM at the start of this work, while OLMo 3 was released after benchmark construction was already underway.

3.2 Training Data of OLMo 2

We summarize OLMo 2's training data as follows: (i) **pre-training** (seven domains): DCLM-Baseline (web), StarCoder (code), peS2o and arXiv (academic), OpenWebMath and Algebraic Stack (math), and Wikipedia/Wikibooks (encyclopedic); (ii) **mid-training** (six domains): three reused from pre-training (filtered DCLM-Baseline, peS2o, and Wikipedia/Wikibooks) and three new domains: FLAN (instruction data), Stack Exchange (Q&A), and Dolmino Math Mix (synthetic math). (iii) **post-training** (three sequential phases): supervised fine-tuning (SFT), direct preference optimization (DPO), and reinforcement learning with verifiable rewards (RLVR). See details in Appendix C.

3.3 Benchmark Overview

As summarized in Table 2, OLMO-DETECT comprises nine domains spanning three training stages of OLMo 2 training data. Mid-training domains reused from pre-training are omitted. Four domains are additionally excluded: Wikipedia/Wikibooks, arXiv, Algebraic Stack, and FLAN, with reasons detailed in the following sections.

For each domain, the contaminated split is explicitly aligned with its uncontaminated counterpart along up to three axes: text quality, temporal range, and lexical similarity. Furthermore, the uncontaminated split is rigorously filtered against the OLMo 2 training data to remove latent contamination.

3.4 Uncontaminated Data Construction

We construct uncontaminated splits in two steps: (i) identifying source-matched candidate pools, and (ii) sampling with 13-gram filtering via infini-gram. Details are provided in Appendix E.

3.4.1 Source-Matched Candidate Pools

To ensure distributional match, we first identify domains for which uncontaminated candidates can be obtained either from held-out data or from the remainder of the original source corpus used to construct OLMo 2's training data.

Held-Out Data. Six domains have held-out sets: (i) pre-training: peS2o, arXiv, OpenWebMath, and Algebraic Stack; (ii) mid-training: Dolmino Math Mix's GSM8K subset; (iii) post-training: RLVR's CoT-GSM8K (8-shot) and CoT-MATH (3-shot).

Remainder of Original Corpora. Six domains are derived from larger upstream corpora: (i) pre-training: DCLM-Baseline and StarCoder; (ii) mid-training: Stack Exchange; and (iii) post-training (Appendix F): SFT, DPO, and the IFEval subset of RLVR. Among these, three domains are obtained through domain-specific quality filtering of original corpora: DCLM-Baseline (fastText scores), StarCoder (GitHub stars and top-1/top-2 word frequency thresholds), and Stack Exchange (question/answer vote counts).

Excluded Domains. Wikipedia/Wikibooks is excluded, as no source-matched data are available.

3.4.2 Sampling with 13-Gram Filtering

Given the source-matched candidate pools, we construct uncontaminated splits via sampling with infini-gram–based 13-gram filtering.

Infini-Gram. Infini-gram (Liu et al., 2024) is a suffix-array-based indexing system supporting efficient n-gram queries over large text corpora for arbitrarily large n. For each of the four OLMo 2 model sizes (1B, 7B, 13B, 32B), we index the full training corpus across all stages (Appendix G).

Stage	Domain	Subsets	Split	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		Uncon	150	217,117	FastText: 0.01809 ± 0.00007	$2018\text{-}01\text{-}19 \pm 930$d	0.05090
			Con	150	217,090	FastText: 0.01811 ± 0.00000	$2018\text{-}01\text{-}17 \pm 933$d	
	peS2o		Uncon	268	217,211	—	$2022\text{-}12\text{-}15 \pm 14$d	0.07321
			Con	268	217,230	—	$2022\text{-}12\text{-}15 \pm 13$d	
	OpenWebMath		Uncon	151	217,183	—	$2019\text{-}10\text{-}12 \pm 853$d	0.07519
			Con	151	217,364	—	$2019\text{-}10\text{-}17 \pm 862$d	
	StarCoder	Java, Assembly	Uncon	365	217,342	GitHub Stars: 159.72 ± 524.54 Top-1 Freq: 0.31007 ± 0.00845 Top-2 Freq: 0.07988 ± 0.03721	—	0.17624
			Con	365	217,267	GitHub Stars: 160.01 ± 524.47 Top-1 Freq: 0.28940 ± 0.00603 Top-2 Freq: 0.05572 ± 0.02283	—	
Mid-training	GSM8K*		Uncon	1,040	217,163	—	—	0.02585
			Con	1,040	217,132	—	—	
	Stack Exchange	Math Stack Exchange, Stack Overflow	Uncon	439	217,141	Question Vote: 3.43 ± 2.09 Answer Vote: 5.11 ± 2.04	$2014\text{-}12\text{-}04 \pm 1326$d	0.05443
			Con	439	217,152	Question Vote: 4.05 ± 2.07 Answer Vote: 6.19 ± 2.07	$2014\text{-}12\text{-}08 \pm 1305$d	
Post-training	SFT†	Aya, WildChat	Uncon	225	217,243	—	—	0.08951
			Con	225	217,013	—	—	
	DPO†		Uncon	113	217,141	—	—	0.05755
			Con	113	217,287	—	—	
	RLVR	CoT-GSM8K, CoT-MATH, IFEval	Uncon	410	217,165	—	—	0.04473
			Con	410	216,992	—	—	

Table 2: Overall per-domain statistics for OLMO-DETECT. For each domain, the contaminated (Con) split is aligned with its uncontaminated (Uncon) counterpart along up to three axes: Text **Quality**, **Temporal** Range, and **Lexical** Similarity. Text Quality is measured using the domain-specific filtering metric adopted in OLMo 2; Temporal Range is reported as mean date $\pm$ standard deviation (std); Lexical Similarity is computed using Jensen–Shannon divergence (JSD), where lower values indicate closer alignment. Text Quality and Temporal Range are controlled only when the corresponding metadata are available for the domain, whereas Lexical Similarity is always controlled. Uncontrolled axes are denoted by "—". The **Subsets** column lists constituent subsets for multi-subset domains; blank entries indicate single-subset domains. *GSM8K is the only Dolmino Math Mix subset and is used to represent the domain. †SFT and DPO are constructed separately for each model size, as their training data differs across sizes; reported values correspond to the 1B model. Each domain is partitioned into dev and test sets with approximately 1:9 sample and token ratios; only total counts are shown. See full statistics in Appendix D.

13-Gram Filtering. Following prior work (Duan et al., 2024; Xu et al., 2025), we apply 13-gram filtering with a threshold of 20%. Specifically, for each input text, we extract all 13-grams S and compute the fraction that appear in the training corpus indexed by infini-gram:

$$\eta = \frac{|\{s \in S : \text{count}(s) > 0\}|}{|S|}.$$

An input is deemed uncontaminated if $\eta < 20\%$.

Random Sampling for Unfiltered Domains. For domains whose candidate pools come from either held-out data or are drawn from upstream corpora without domain-specific quality filtering, we randomly sample candidates and retain only those passing the 13-gram filter.

Boundary Sampling for Quality-Filtered Domains. For quality-filtered domains (DCLM-Baseline, StarCoder and Stack Exchange), random sampling from the source-corpus remainder would yield data of systematically lower quality than that used to train OLMo 2. We thus apply **boundary sampling**, which samples instances whose quality scores lie immediately below OLMo 2's original filtering threshold, subject to the 13-gram filter.

Token Count Per Domain. To ensure balanced evaluation signal and avoid biasing toward any specific domain, we set a uniform target budget of approximately 217K tokens per domain, determined by the uncontaminated token count in GSM8K.

Domains with Multiple Subsets. For domains spanning multiple sampleable subsets, we select two subsets, except RLVR for which we use all three. Details are provided in Appendix H.

Excluded Domains. ArXiv and Algebraic Stack are excluded as no uncontaminated data are found.

3.5 Contaminated Data Construction

Given the uncontaminated splits, we construct contaminated splits in two steps: (i) identifying candidate pools from OLMo 2's training data, and (ii)

sampling via simulated annealing.

Candidate Pools. For domains without domain-specific quality filters, candidate pools are the corresponding OLMo 2 training data. For quality-filtered domains, we apply **boundary sampling**, restricting candidates to instances whose quality scores lie immediately above OLMo 2's filtering threshold, yielding candidates closely matched in quality to their uncontaminated counterparts.

Alignment Axes. Beyond **text quality** alignment enforced through boundary sampling, we additionally align contaminated and uncontaminated splits along two axes: (i) **temporal range**, since temporal drift can artificially inflate detection performance (Duan et al., 2024); and (ii) **lexical similarity**, measured by Jensen–Shannon divergence (JSD) between their token-frequency distributions, since verbatim contamination is fundamentally a surface-form phenomenon and lexical mismatch would directly confound the contamination signal. Temporal alignment is applied only when metadata are available; lexical alignment is always enforced.

Constrained Sampling via Simulated Annealing. Given the candidate pools, we perform constrained sampling to select contaminated instances matching the uncontaminated splits in both instance (exact) and token counts (within ± 250), thus eliminating average instance length as a confounder, while minimizing a distributional loss $\mathcal{L}$ via simulated annealing (Lecchini-Visintini et al., 2007), an optimization technique with probabilistic acceptance of suboptimal moves to escape local minima:

$$\mathcal{L} = \mathrm{JSD}(p, q) + \mathbb{1}[\text{temporal metadata}] \cdot \mathcal{L}_{\text{temp}},$$

where p and q denote the contaminated and uncontaminated token distributions, JSD enforces lexical alignment, and $\mathcal{L}_{\text{temp}}$ aligns the mean and std of dates between uncontaminated and contaminated splits (applied only when metadata are available). Details are provided in Appendix I.

Excluded Domains. FLAN is excluded, as contaminated data differ substantially in average tokens per instance from uncontaminated data.

3.6 Decontamination Validation

Table 3 shows that the uncontaminated and contaminated splits of OLMo-Detect share an average 7-gram overlap of around 1.6% across all domains, with the maximum reaching just 6.5%, confirming the effectiveness of the 13-gram filtering.

Stage	Domain	Overlap
Pre-training	DCLM	0.00117
	OpenWebMath	0.00242
	peS2o	0.00193
	StarCoder	0.01574
Mid-training	Math-Mix	0.06502
	Stack Exchange	0.00402
Post-training	SFT	0.01226
	DPO	0.01050
	RLVR	0.03109
Avg		**0.01602**

Table 3: Average 7-gram overlap between contaminated ($\mathcal{C}$) and uncontaminated ($\mathcal{U}$) splits in OLMo-Detect, computed as $|\mathcal{C} \cap \mathcal{U}|/|\mathcal{C}|$. Per-domain values are averaged across all model sizes. The bottom **Avg** row reports the overall average across all domains and model sizes. Per-model breakdown is provided in Appendix J.

3.7 OLMo-Detect (Shifted)

To examine the effect of distributional shift on detection robustness, we introduce OLMo-Detect (Shifted), a variant of OLMo-Detect where distributional mismatches are intentionally permitted. Mid-training/GSM8K and post-training/RLVR are excluded from OLMo-Detect (Shifted), as random sampling from their training data does not produce contaminated splits with substantially different JSD from the uncontaminated counterparts. Across the remaining seven domains, uncontaminated splits are unchanged from OLMo-Detect, while contaminated splits are constructed by randomly sampling from the corresponding OLMo 2 training data to match only the instance count and token count of the uncontaminated counterparts, without distributional alignment. Full statistics are provided in Appendix K.

4 Experimental Settings

Models. We evaluate the four instruction-tuned OLMo 2 models: 1B, 7B, 13B, and 32B.

Compared Methods. We evaluate 15 state-of-the-art methods spanning four families: (i) likelihood-based: Perplexity (PPL) (Yeom et al., 2018), Zlib and Lowercase (Lower) (Carlini et al., 2021), Neighborhood Attack (Neigh.) (Mattern et al., 2023), MinK (Shi et al., 2023), MinK++ (Zhang et al., 2024a), DCPDD (Zhang et al., 2024b), PAC (Ye et al., 2024), RECALL (Xie et al., 2024), CAMIA (Chang et al., 2025), and EMMIA

Model	Stage	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
1B	Pre	0.6321	0.6795	0.6650	0.5594	0.5423	0.6034	0.5784	**0.7017**	0.6738	0.5964	0.6387	0.5916	0.5430	0.5616	0.6119
	Mid	0.6327	0.6588	0.6650	0.6555	0.5494	0.6602	0.3609	**0.7130**	0.6706	0.5986	0.5607	0.4790	0.5131	0.4801	0.5855
	Post	0.5232	0.5570	0.5668	0.5128	0.5158	0.5025	0.4950	0.5820	**0.5835**	0.5146	0.5630	0.4842	0.5682	0.4894	0.5327
	Overall	0.5779	0.6087	0.6312	0.5833	0.5256	0.5926	0.4614	**0.6701**	0.6281	0.5651	0.5961	0.5084	0.5310	0.5028	0.5702
7B	Pre	0.6477	**0.7126**	0.6595	0.6084	0.5552	0.7080	0.5690	0.7113	0.6910	0.5972	0.6073	0.5454	0.5523	0.5665	0.6237
	Mid	0.7062	0.7433	0.7599	0.7650	0.5412	0.8020	0.2491	**0.8119**	0.7268	0.6558	0.5558	0.4964	0.5345	0.5044	0.6323
	Post	0.5549	0.5606	0.5672	0.5162	0.4992	0.5149	0.6103	**0.7027**	0.5746	0.5229	0.5567	0.4658	0.5251	0.4816	0.5466
	Overall	0.6257	0.6591	0.6791	0.6476	0.5322	0.6772	0.4554	**0.7564**	0.6462	0.5870	0.5790	0.5045	0.5343	0.5154	0.5999
13B	Pre	0.6508	**0.7211**	0.6698	0.6242	0.5620	0.6955	0.5923	0.6993	0.7042	0.6027	0.6171	0.5521	0.5806	0.5856	0.6327
	Mid	0.8156	0.8607	0.8488	0.9138	0.5765	**0.9315**	0.1565	0.9011	0.8138	0.7670	0.5994	0.4995	0.5734	0.5241	0.6987
	Post	0.5504	0.5736	0.5937	0.5211	0.5187	0.5141	0.5166	**0.6612**	0.5934	0.5328	0.4676	0.4414	0.5582	0.4679	0.5365
	Overall	0.6786	0.7250	0.7280	0.7330	0.5502	0.7480	0.3729	**0.7799**	0.6784	0.6285	0.5633	0.5017	0.5548	0.5278	0.6264
32B	Pre	0.6675	**0.7302**	0.6705	0.6559	0.5739	0.7153	0.6653	0.6902	0.7071	0.6013	0.6607	0.4774	0.5272	0.5985	0.6386
	Mid	0.7808	0.8042	0.8220	0.8411	0.5412	0.8787	0.2684	**0.9020**	0.8221	0.7917	0.6278	0.5754	0.6124	0.5328	0.7000
	Post	0.5440	0.5712	0.5721	0.5306	0.5113	0.5432	0.5674	**0.6325**	0.5877	0.5307	0.4488	0.4991	0.5571	0.4555	0.5394
	Overall	0.6500	0.6803	0.6929	0.6810	0.5427	0.7171	0.4784	**0.7683**	0.6965	0.6356	0.5827	0.5221	0.5538	0.5316	0.6238
Model-Avg	Pre	0.6495	**0.7108**	0.6662	0.6120	0.5584	0.6805	0.6013	0.7006	0.6940	0.5994	0.6309	0.5416	0.5508	0.5780	0.6267
	Mid	0.7338	0.7668	0.7739	0.7938	0.5521	0.8181	0.2587	**0.8320**	0.7583	0.7033	0.5859	0.5126	0.5584	0.5103	0.6541
	Post	0.5431	0.5656	0.5750	0.5202	0.5112	0.5187	0.5473	**0.6446**	0.5848	0.5252	0.5090	0.4726	0.5522	0.4736	0.5388
	Overall	0.6331	0.6683	0.6828	0.6612	0.5377	0.6837	0.4420	**0.7437**	0.6623	0.6040	0.5803	0.5092	0.5435	0.5194	0.6051

Table 4: Experimental results on OLMO-DETECT, reported using AUC. **Pert.** denotes perturbation-based, while **Out.-Dist.** denotes output-distribution-based. **Pre/Mid/Post** report stage-wise AUC for pre-training, mid-training, and post-training, while ***Overall*** reports overall AUC by pooling contaminated and uncontaminated instances across all stages. **Model-Avg** averages across all model sizes, while **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). Per-domain AUC results are provided in Appendix M.

Domain	1B	7B	13B	32B
DPO	0.4302	0.4882	0.4206	0.5054
RLVR	0.5051	0.4392	0.4392	0.4206

Table 5: AUC of SelfCrit on DPO and RLVR.

(Kim et al., 2026); (ii) prompt-based: Guided Instruction (Guided) (Golchin and Surdeanu, 2023); (iii) perturbation-based: DCQ (Golchin and Surdeanu, 2025); and (iv) output-distribution-based: CDD (Dong et al., 2024) and Self-Critique (SelfCrit) (Tao et al., 2025) (proposed for RL and thus evaluated only on DPO and RLVR). Implementation details are provided in Appendix L.

Metrics. Following prior work (Shi et al., 2023), we use AUC and TPR@5%FPR. For DCQ and Guided, whose outputs are categorical, we extract continuous per-sample scores (Appendix L).

5 Experimental Results

We aim to investigate the following research questions (RQs). On OLMO-DETECT, **RQ1**: How do existing detection methods perform? **RQ2**: What domain characteristics are associated with detection difficulty? **RQ3** and **RQ4**: How does detection performance vary across stages and model sizes? On OLMO-DETECT (SHIFTED), **RQ5**: How robust are existing methods to distributional shift?

5.1 RQ1: Overall Performance

Table 4 reports the AUC of existing methods on OLMO-DETECT; TPR@5%FPR results are provided in Appendix N. Overall, no method achieves *strong* performance (AUC $\geq$ 0.75), with the best-performing method CAMIA reaching only 0.7437 on average. This overall limited performance suggests that, under distribution-matched settings with minimal confounding factors, genuine contamination signals become substantially harder to detect.

Across method families, likelihood-based methods generally outperform other categories, with two exceptions: (i) MinK++, which applies token-level vocabulary normalization to MinK yet performs substantially worse, suggesting that fine-grained normalization may weaken the detection signal under distribution-matched settings; and (ii) RECALL, which assumes that non-member prefixes degrade member predictions, inverts on mid-training/GSM8K (Table 34), as prefixes from the dev set act as in-context demonstrations, boosting member predictions instead of disrupting them.

As shown in Table 5, SelfCrit achieves near-random AUC on both DPO and RLVR across all model sizes, in stark contrast to the original paper's reported score on RL-MIA (Tao et al., 2025), a

Stage	Domain	AUC	Count	Framing
Pre	DCLM	0.5645	9.00	✗
	OpenWebMath	0.5382	1.53	✗
	peS2o	0.5309	2.02	✗
	StarCoder/Java	0.7521	1.00	✓
Mid	GSM8K	0.6958	6.00	✓
	Stack Exchange	0.6130	6.00	✗
Post	SFT	0.5865	2.34	✗
	DPO	0.5039	1.10	✗
	RLVR	0.5550	1.19	✗

Table 6: Domain-level AUC (averaged across all methods and model sizes), duplication **count** (averaged across model sizes), and domain-specific structural **framing**. StarCoder/Assembly is omitted as its small sample size may confound the analysis (Appendix O).

Subset	Setting	Avg AUC
CoT-GSM8K	w 8-shot CoT	0.7211
	w/o 8-shot CoT	0.7585
CoT-MATH	w 3-shot CoT	0.6637
	w/o 3-shot CoT	0.6684

Table 7: Effect of removing the shared boilerplate CoT prefix from RLVR's GSM8K and MATH on AUC.

benchmark constructed via controlled, isolated RL training on a curated corpus. This gap indicates that contamination detection in realistic multi-stage post-training pipelines is substantially harder than in simulated single-stage RL setups.

5.2 RQ2: Domain-Level Detectability

To tackle RQ2, we examine three candidate factors: (i) duplication count, which captures repeated exposure; (ii) domain-specific structural framing, a qualitative indicator of whether samples follow recurring formats unique to a given domain that wrap sample-specific content; and (iii) training stage, which reflects both training objective and data introduction order. Our findings are as follows. (i) **Domain-specific structural framing is associated with higher detectability.** As shown in Figure 1, there are two domains with structural framing: StarCoder/Java (keywords) and GSM8K (Question/Reasoning/Answer template). Table 6 shows that domains with framing achieve AUC above 0.69, whereas domains without it (Appendix O) fall below 0.62. A controlled comparison is GSM8K vs. Stack Exchange: both are mid-training domains with the same duplication count and training objective, yet GSM8K, which wraps unique content in a Question/Reasoning/Answer template, reaches

(a) Pre-training/StarCoder/Java, Avg AUC 0.75

```
package org.bc.jcajce.provider.asymmetric;
import org.bc.asn1.eac.EACObjectIdentifiers;
import org.bc.asn1.x9.X9ObjectIdentifiers;
public class EC { ... }
```

(b) Mid-training/GSM8K, Avg AUC 0.70

```
Question: In Hawaii they get 2 in/day...
Reasoning: They need 730 inches a year...
Answer: 3
```

Figure 1: Representative samples illustrating domain-specific structural framing. Recurring framing tokens consistent across samples are highlighted in gray; unique content varying per sample is in yellow.

0.70 AUC, while Stack Exchange, which lacks consistent markers, reaches 0.61. We hypothesize that stable, domain-specific structural framing may help anchor sample-specific content, making memorized instances easier to detect.[1] However, recurring verbatim boilerplate does not aid detection: as illustrated in Table 7, removing the CoT prefix from CoT-GSM8K and CoT-MATH yields higher AUCs, suggesting that boilerplate can dilute the detection signal. (ii) **Duplication count and training stage alone do not explain detectability.** Table 6 exhibits a weak correlation between duplication count and detection AUC (Pearson r = 0.14): StarCoder/Java has the lowest count but the highest AUC, DCLM has the highest count but only 0.56 AUC, and GSM8K and Stack Exchange share the same count yet differ by 0.09 AUC. Training stage also shows no simple pattern. Among domains with similar counts and no structural framing, the ordering reverses across stages: SFT (0.59) exceeds peS2o (0.53), whereas DPO (0.50) falls below OpenWebMath (0.54), ruling out a monotonic stage effect. We suspect that both factors likely contribute alongside structural framing, but their standalone effects are difficult to isolate.

5.3 RQ3: Cross-Stage Performance

As shown in Table 4, most methods exhibit a non-monotonic pattern across stages: performance peaks at mid-training, followed by pre-training, and is weakest at post-training. However, this ordering largely reflects each stage's domain-specific structural framing (RQ2) rather than an independent stage effect (Table 6): mid-training is lifted

[1] Chat-template tokens (<|user|>, <|assistant|>) are not considered domain-specific framing, as they appear identically across SFT, DPO, and RLVR. See Appendix O.

Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
1B	Matched	0.5835	0.6325	0.6083	0.5403	0.5516	0.5602	0.5527	0.6431	0.6253	0.5668	0.6143	0.5335	0.5394	0.5192	0.5765
	Shifted	0.6689	0.6324	0.5648	0.9980	0.2781	1.0000	0.5953	0.6752	0.7111	0.6299	0.6779	0.4856	0.5597	0.5105	0.6420
7B	Matched	0.6051	0.6613	0.6171	0.5834	0.5689	0.6306	0.5985	0.6717	0.6407	0.5738	0.5974	0.5199	0.5361	0.5088	0.5938
	Shifted	0.6666	0.6387	0.5870	0.9702	0.5605	0.9963	0.6136	0.6898	0.7751	0.5083	0.5513	0.5180	0.6478	0.5017	0.6589
13B	Matched	0.6046	0.6681	0.6322	0.5946	0.5830	0.6295	0.6328	0.6757	0.6491	0.5758	0.5693	0.5286	0.5659	0.5369	0.6033
	Shifted	0.6692	0.6432	0.5805	0.9159	0.6026	1.0000	0.6209	0.6733	0.7293	0.6237	0.6544	0.5559	0.6122	0.5612	0.6745
32B	Matched	0.6169	0.6700	0.6211	0.6164	0.5787	0.6490	0.6781	0.6579	0.6587	0.5753	0.5979	0.5279	0.5318	0.5395	0.6085
	Shifted	0.6871	0.6537	0.5662	0.9313	0.5839	0.9979	0.7179	0.6858	0.7111	0.6271	0.6207	0.5100	0.5330	0.5634	0.6707
Avg Δ		+0.0704	−0.0160	−0.0450	+0.3702	−0.0643	+0.3812	+0.0214	+0.0189	+0.0882	+0.0243	+0.0313	−0.0101	+0.0449	+0.0081	+0.0660
***p*-value**		<0.01	0.06	<0.01	<0.01	0.43	<0.01	0.19	0.09	0.01	0.48	0.36	0.57	0.16	0.45	<0.01
Class.		I	D	D	I	R	I	R	I	I	R	R	R	R	R	I

Table 8: Comparison between OLMO-DETECT (**Matched**) and OLMO-DETECT (SHIFTED) (**Shifted**), reported using **all-stage** AUC. AUC cells are colored by magnitude: weak (<0.60), moderate (0.60–0.75), strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df = 3). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$. Detailed results are shown in Appendix P.

by GSM8K; pre-training contains StarCoder/Java alongside three free-form domains; and post-training contains no framed domain.

5.4 RQ4: Cross-Scale Performance

The **Method-Avg** column of Table 4 shows that average detection AUC improves from 0.57 at 1B to 0.63 at 13B, then plateaus (0.62 at 32B). The saturation appears differently across stages: pre-training gains slowly with scale (0.61 to 0.64), mid-training rises to 0.70 at 13B then plateaus, and post-training is essentially scale-invariant (∼0.54). This suggests that once a model is sufficiently large (around 13B in OLMo 2), further scaling yields limited additional detection benefit. This finding appears to contradict prior work showing that verbatim memorization itself scales log-linearly with model size (Carlini et al., 2022). However, the two findings are not contradictory: memorization and detectability need not scale identically. Even if memorization continues to grow at 32B, current methods may recover only a bounded fraction of that signal, especially under our distribution-matched setting where confounders are minimized.

5.5 RQ5: Robustness to Distribution Shift

As illustrated in Table 8, existing methods exhibit varying robustness to distributional shift: 7 methods are *robust*, 5 are *inflated*, and 2 are *degraded*. Our key findings are as follows. (i) **Methods relying on raw or distribution-sensitive token likelihoods inflate the most.** MinK and DCPDD reach AUC > 0.9 on OLMO-DETECT (SHIFTED), as token-level likelihoods are systematically lower for out-of-distribution non-members, creating an artificially large member-vs-non-member gap that reflects distributional mismatch. (ii) **Degraded methods use distribution-sensitive normalizers.** Zlib and Lower normalize PPL by compressed length and lowercased PPL, respectively, aiming to factor out text complexity. Under distributional shift, however, these normalizers can react more strongly to surface-form changes than the raw loss, partially canceling the membership signal and shrinking the gap that raw-likelihood methods amplify. MinK++, although classified as *robust* overall, shows the same instability at 1B; its per-token z-score normalization stabilizes only at larger scales where the model becomes better calibrated. (iii) **Robustness does not imply strong detection.** Several *robust* methods (e.g., DCQ and Guided) remain near chance in both matched and shifted settings. An ideal method would achieve *strong* performance on OLMO-DETECT and be classified as *robust* on OLMO-DETECT (SHIFTED); unfortunately, no current method achieves both.

6 Conclusion

We introduced OLMO-DETECT, the first benchmark for verbatim contamination detection in LLMs covering all three stages of the modern training pipeline, with contaminated and uncontaminated splits explicitly aligned along multiple axes and rigorously verified. We also introduced OLMO-DETECT (SHIFTED), a complementary variant that assesses method robustness under distributional shift. Evaluating 15 state-of-the-art methods, we find that no current method is both *strong* on OLMO-DETECT and *robust* on OLMO-DETECT (SHIFTED), highlighting the unresolved challenge of detecting genuine contamination signals once distributional confounders are removed.

Limitations

Reliance on the OLMo 2 series. Our benchmark spans only the OLMo 2 model family (1B, 7B, 13B, and 32B). This choice is unavoidable for scientific rigor: at the time we began benchmark construction, OLMo 2 was the only model family providing full access to the pre-training, mid-training, and post-training data required to build a reliable multi-stage verbatim-contamination benchmark. While our experiments are based on OLMo 2, an increasing number of modern LLMs follow a similar three-stage training pipeline (Grattafiori et al., 2024; DeepSeek-AI et al., 2025; Abdin et al., 2024; Yang et al., 2025), suggesting that our findings may generalize beyond OLMo 2. We hope more model developers will release full training data, enabling direct extension of our methodology to broader model families.

Use of OLMo 2 rather than OLMo 3. At the time of our benchmark construction, OLMo 2 was the most recent fully open-source LLM family; OLMo 3 was released concurrently with this work, after construction on OLMo 2 was already underway. Beyond this timing consideration, OLMo 2 offers two methodological advantages: (i) it spans four model sizes (1B, 7B, 13B, and 32B), whereas OLMo 3 provides only two (7B and 32B), enabling finer-grained analysis of how detection performance scales with model size; and (ii) its post-training follows a linear pipeline, whereas OLMo 3 splits post-training into multiple parallel branches, making OLMo 2 better suited for systematic phase-by-phase analysis. Notably, our benchmark construction pipeline is generic and directly applicable to OLMo 3; we leave this extension for future work.

Inability to isolate duplication count and training-stage effects on detectability. Our analysis in Section 5.2 identifies domain-specific structural framing as the primary predictor of detectability, but we cannot isolate the residual effects of duplication count and training stage. Investigating standalone effects would require a controlled study with curated duplication schedules and matched domain compositions. This is infeasible in our setting: we cannot modify duplication counts retroactively in OLMo 2's fixed training corpus, and varying these factors via additional fine-tuning would deviate from the actual multi-stage LLM training pipeline, producing unreliable results that no longer reflect realistic contamination behavior.

Focus on verbatim contamination. Our benchmark targets verbatim contamination, where an evaluation sample appears in an LLM's training data without modification. This is the most well-studied and operationally relevant form of contamination, as it can directly inflate benchmark scores (Sainz et al., 2023). However, more complex forms, such as paraphrased or back-translated contamination, are not covered by our benchmark or by most of the detection methods we evaluate. Extending our framework to non-verbatim contamination is an important direction for future work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku.

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

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. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv: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. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information*

Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

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. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

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

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

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, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jasper Dekoninck, Mark N Müller, and Martin Vechev. 2024. Constat: Performance-based contamination detection in large language models. *Advances in Neural Information Processing Systems*, 37:92420–92464.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. 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)*, pages 8706–8719.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

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

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

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. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

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

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv: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, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

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, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *Preprint*, arXiv:2506.04178.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. 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*, pages 427–431. Association for Computational Linguistics.

Gyuwan Kim, Yang Li, Evangelia Spiliopoulou, Jie Ma, and William Yang Wang. 2026. 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)*, pages 1115–1129.

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, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

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

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

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, and 40 others. 2025. Datacomp-lm: In search of the next generation of training sets for language models. *Preprint*, arXiv: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, and 48 others. 2023. Starcoder: may the source be with you! *Preprint*, arXiv:2305.06161.

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

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

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

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, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

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, and 24 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

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, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and

262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

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. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv: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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. 2025. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5058–5070.

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

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

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

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

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, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

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

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

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

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. 2024. Redpajama: an open dataset for training large language models. *Preprint*, arXiv: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. 2021. Finetuned lan-

guage models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

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. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *Preprint*, arXiv:2203.05482.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2025. 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*, pages 2742–2785.

Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2024. Recall: Membership inference via relative conditional log-likelihoods. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8671–8689.

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A Smith, and Hannaneh Hajishirzi. 2025. 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*, pages 24955–24980.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

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

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024b. 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*, pages 5263–5274.

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

A Detailed Descriptions of Contamination Detection Benchmarks

We provide detailed descriptions of each benchmark surveyed in Section 2.

WikiMIA (Shi et al., 2023). A temporal-cutoff benchmark constructed from Wikipedia event pages. Articles describing events before January 2017 are treated as contaminated, while those referring to events after 2023 are treated as uncontaminated. The benchmark is partitioned by length into subsets of 32, 64, and 128 tokens.

StackMIA (Ye et al., 2024). A temporal-cutoff benchmark constructed from Stack Exchange posts. Posts dated before January 2017 are treated as contaminated, while posts dated after May 2023 are treated as uncontaminated.

DETCON (Dong et al., 2024). A model-output-based benchmark designed to simulate contamination during SFT. A base LLM is fine-tuned via low-rank adaptation (LoRA) (Hu et al., 2021) on HumanEval (Chen et al., 2021) (code generation) and GSM8K (mathematical reasoning) under varying mixing ratios, learning rates, and occurrence counts. Outputs generated by the fine-tuned model are treated as contaminated, while outputs generated by the corresponding base model are treated as uncontaminated.

MIMIR (Duan et al., 2024). A partition-based benchmark constructed from The Pile (Gao et al., 2020) and Dolma (Soldaini et al., 2024), where training instances are treated as contaminated and held-out instances from the same domains are treated as uncontaminated. MIMIR spans seven domains from The Pile (Wikipedia, GitHub, Pile-CC, PubMed Central, ArXiv, DM Mathematics, HackerNews) and five domains from Dolma (Wikipedia, C4, Reddit, Common Crawl, peS2o). A 13-gram Bloom filter is used to remove near-duplicates between contaminated and uncontaminated splits, though substantial 7-gram overlap remains in several domains. Specifically, average 7-gram overlap reaches 32.5%, 39.3%, and 41.0% in Wikipedia, ArXiv, and PubMed Central, respectively, and 76.9% and 72.8% in GitHub and DM Mathematics.

PatentMIA (Zhang et al., 2024b). A temporal-cutoff benchmark built from Chinese-language Google Patents pages. Patents published before January 2023 are treated as contaminated, while those with priority dates after March 2024 are treated as uncontaminated. Each instance consists of a 512-word snippet randomly extracted from the corresponding page.

WikiMIA-24 (Fu et al., 2025). A temporal-cutoff benchmark that updates WikiMIA's cutoff to evaluate more recent LLMs. Articles before January 2017 remain as contaminated, while articles after March 2024 are treated as uncontaminated.

RL-MIA (Tao et al., 2025). A partition-based benchmark designed to simulate contamination during RL post-training. RL-MIA draws from four source benchmarks: AIME24, AIME25, K&K (Xie et al., 2025), and SAT (Liu et al., 2025), with GSM8K additionally used as a case study to disentangle pre-training and RL-stage contamination. For AIME24 and AIME25, 50% of each benchmark's items are injected into the OpenR1-Math-46K (Guha et al., 2025) RL training corpus as contaminated samples, while the remaining items serve as uncontaminated samples. For K&K and SAT, original training instances are treated as contaminated, while additional items synthesized using the same generation procedure are treated as uncontaminated. For GSM8K, half of the test set is injected during RL training as contaminated, while the remaining half is held out as uncontaminated.

OLMoMIA (Kim et al., 2026). A partition-based benchmark constructed from intermediate checkpoints of OLMo 1-7B (Groeneveld et al., 2024) and its pre-training corpus Dolma. Instances seen before training step 100k are treated as contaminated, while instances introduced between steps 400k and 500k are treated as uncontaminated. To control distributional overlap, instances are embedded using NV-Embed-v2 (Lee et al., 2025) and clustered via K-means with $K = 50$. Six difficulty variants (Easy, Medium, Hard, Random, Mix-1, Mix-2) are then constructed by selecting cluster pairs and within-cluster examples based on inter-cluster distance, ranging from maximum separability (Easy) to minimum separability (Hard). Each variant contains two length-based subsets (64 and 128 tokens), with 500 contaminated and 500 uncontaminated examples per subset.

Distributional Shift. Existing benchmarks are constructed in ways that may reflect distributional shift rather than genuine contamination signals: (1) **temporal cutoffs** (WikiMIA, StackMIA, PatentMIA, WikiMIA-24) that introduce chronological shifts; (2) **train/held-out partitioning** (MIMIR, RL-MIA) without explicit control over distributional alignment; and (3) **model-output partitioning** (DETCON) between a fine-tuned model and its base version, introducing distributional differences from the fine-tuning procedure. OLMoMIA is the only exception, aligning along semantic similarity. However, semantic similarity may not be ideal for verbatim contamination detection, as the task fundamentally concerns surface-form overlap rather than semantic equivalence.

B Taxonomy of Contamination Detection Methods

This section provides a taxonomy of existing methods for verbatim contamination detection in LLMs.

Likelihood-based. These methods infer contamination from the model's token-level likelihoods. Perplexity (Yeom et al., 2018) uses the raw negative log-likelihood. Zlib and Lowercase (Carlini et al., 2021) compute likelihood ratios against the input's compressed length and lowercased-version likelihood, respectively. Neighborhood (Mattern et al., 2023) compares likelihoods against synthetic neighbors. MinK (Shi et al., 2023) and its extension MinK++ (Zhang et al., 2024a) aggregate over the least-probable tokens, with the latter additionally normalizing each token's log-probability using vocabulary-level statistics. DCPDD (Zhang et al., 2024b) calibrates likelihood against a corpus-level token-frequency prior, and PAC (Ye et al., 2024) compares likelihoods against polarized input perturbations. RECALL (Xie et al., 2024) compares the candidate's conditional likelihood under a non-member prefix against its unconditional likelihood. CAMIA (Chang et al., 2025) derives context-aware signals from the per-token loss sequence, combining prefix-property calibration with loss-dynamics statistics. EMMIA (Kim et al., 2026) extends RECALL by iteratively refining prefix selection via expectation-maximization. Beyond these reference-free methods, Smaller Ref (Carlini et al., 2021) and SPV-MIA (Fu et al., 2024) additionally incorporate a separately trained reference model.

Ordering-based. These methods exploit the model's preference for the canonical ordering of training data. Permutation Test (Oren et al., 2023) compares model likelihoods under canonical and shuffled benchmark orderings at the dataset level, whereas Local Order Quiz (Samuel et al., 2025) probes the same signal at the instance level.

Prompt-based. These methods elicit memorization through constructed prompts. Guided Instruction (Golchin and Surdeanu, 2023) prompts the model with the dataset name, partition, and an initial segment to flag exact or near-exact completions of the remaining text. Chang et al. (2023) use a name-cloze formulation that masks a proper name, while TS-Guessing (Deng et al., 2024) masks an option in a multiple-choice question.

Perturbation-based. DCQ (Golchin and Surdeanu, 2025) casts contamination detection as a multiple-choice task where the model is asked to identify the original candidate among several paraphrased variants; contamination is flagged when the model consistently selects the original text.

Output-distribution-based. These methods analyze statistical properties of model outputs. CDD (Dong et al., 2024) measures the peakedness of the model's output distribution, while Self-Critique (Tao et al., 2025) detects RL-phase contamination through output entropy collapse.

Performance-based. ConStat (Dekoninck et al., 2024) compares the model's task performance against its expected performance, estimated from a reference benchmark of comparable difficulty, and flags contamination when the gap is significant.

C Detailed Descriptions of OLMo 2 Training Data

This appendix provides a detailed breakdown of OLMo 2's training data across the three stages. Token counts, document counts, and filtering details are reproduced from 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 DCLM-Pool filtered using a fastText classifier (Joulin et al., 2017). From Dolma 1.7 (Soldaini et al., 2024), it incorporates five subsets: peS2o (Soldaini and 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 is summarized in Table 9. The 1B and 7B models pretrain on 4 trillion tokens of this mix (1 epoch), the 13B on 5 trillion tokens (1.2 epochs), and the 32B on 6 trillion tokens (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 is a heterogeneous mixture aggregating synthetic math data (TuluMath, generated with GPT-4o (OpenAI et al., 2024a); 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 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: 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 ≥ 2 ($\sim 20.3\%$ retained overall). FLAN is decontaminated by removing train-

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).

ing, 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).

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 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[2]; 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)

[2]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.

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.

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.

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).

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%).

D Detailed Per-Domain Statistics of OLMo-Detect

We provide detailed per-domain statistics for OLMo-Detect, as introduced in Section 3.3. Per-domain statistics for the 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 (for hyperparameter tuning) and a test set (for method evaluation), with both sample and token ratios set to approximately 1:9.

A few observations warrant brief explanation. (i) **Text quality.** For quality-controlled domains, uncontaminated splits have slightly lower quality scores than their contaminated counterparts (e.g., DCLM-Baseline's FastText: 0.01809 vs. 0.01811). This is a direct consequence of boundary sampling: uncontaminated instances are drawn from just below OLMo 2's filter threshold, while contaminated instances are drawn from just above. This is the tightest quality alignment achievable, since any closer matching would require uncontaminated instances to be in the training data. (ii) **Lexical similarity.** For the majority of domains, we achieve a JSD below 0.15, indicating strong lexical alignment. The sole exception is StarCoder/Java (JSD = 0.19607), where the high lexical diversity of Java code (e.g., project-specific class names, variable identifiers, and import paths) produces token distributions that vary substantially across files. This residual JSD is data-driven rather than a methodological limitation, as confirmed by the contrasting low JSD of StarCoder/Assembly (0.01720) under the same construction pipeline.

E Detailed Uncontaminated and Contaminated Data Construction

This appendix details the construction of uncontaminated and contaminated splits for each domain in OLMo-Detect. Across all domains, contaminated splits are generated via constrained sampling with simulated annealing to match their uncontaminated counterparts in instance count (exact), token count (within ±250 tokens), and applicable alignment axes (temporal range when metadata are available and lexical similarity in all cases). We employ simulated annealing because subset selection is a combinatorial optimization problem with a non-convex objective, making gradient-free stochastic search well suited for escaping local minima. A summary is illustrated in Table 22.

Pre-training/DCLM-Baseline. Uncontaminated. Candidates are drawn from three shards of the upstream DCLM-Pool corpus: `global-shard_01-local-shard_0`, `global-shard_05-local-shard_6`, and `global-shard_07-local-shard_3`. We identify the minimum fastText score in OLMo 2's DCLM-Baseline (0.01811) and retain instances with scores in (0.01800, 0.01811) that pass the 13-gram filter. **Contaminated.** Candidates are drawn from the corresponding DCLM-Baseline shards and restricted to instances with a fastText score of 0.01811.

Pre-training/peS2o. Uncontaminated. Candidates are randomly drawn from peS2o's two held-out files (`validation-00.jsonl` and `validation-01.jsonl`), subject to the 13-gram filter. **Contaminated.** Candidates are drawn from OLMo 2's peS2o training data.

Pre-training/OpenWebMath. Uncontaminated. Candidates are randomly drawn from OpenWebMath's held-out file (`test.jsonl`), subject to the 13-gram filter. **Contaminated.** Candidates are drawn from OLMo 2's OpenWebMath training data.

Pre-training/StarCoder. We use the Java and Assembly subsets. **Uncontaminated.** Candidates are drawn from the original StarCoder dataset (Li et al., 2023) (Assembly: `train-00.jsonl` and `train-01.jsonl`; Java: `train-00.jsonl` through `train-86.jsonl`). After retaining repositories with at least 2 GitHub stars (matching OLMo 2's repository-level filter), we apply boundary sampling: candidates are restricted to instances that narrowly fail OLMo 2's word-frequency filter, where exactly one of the two thresholds (top-1 word frequency at most 0.30, combined top-1 plus top-2 frequency at most 0.50) is exceeded by less than 0.01 while the other is satisfied. The 13-gram filter is then applied. **Contaminated.** Candidates are drawn from OLMo 2's StarCoder training data (Assembly: `assembly-00.jsonl` and `assembly-01.jsonl`; Java: `00.jsonl` through `86.jsonl`), restricted to instances whose top-1 frequency is at least 0.28 or combined frequency is at least 0.48, placing them within 0.02 of OLMo 2's filter cutoffs from the satisfied side. The ≥ 2 GitHub stars constraint is inherited from the training data, ensuring star-count alignment.

Mid-training/Dolmino Math Mix. We sample from the GSM8K subset, which is the only sampleable subset of Dolmino Math Mix (Appendix G). **Uncontaminated.** Candidates are drawn from `gsm8k_test.jsonl`, excluding the 200 instances designated by OLMo et al. (2025) as a development set for evaluation, then passed through the 13-gram filter. The resulting $\sim$217K-token budget establishes the per-domain token budget for all other domains in OLMo-Detect. **Contaminated.** Candidates are drawn from `gsm8k_train.jsonl`.

Mid-training/Stack Exchange. We use the Stack Overflow and Math Stack Exchange forums. **Uncontaminated.** Candidates are drawn from the original 2024-09-30 Stack Exchange dump (after applying OLMo 2's preprocessing pipeline to obtain the same Q&A format), restricted to instances satisfying either (a) question score $= 2$ and answer score ≥ 5, or (b) question score ≥ 3 and answer score $= 4$, subject to the 13-gram filter. **Contaminated.** Candidates are drawn from OLMo 2's Stack Exchange training data (`stackexchange-00.jsonl` through `stackexchange-15.jsonl`), filtered to the same two forums and further restricted to instances with question score $= 3$ or answer score $= 5$.

Post-training/SFT. We use the WildChat and Aya subsets. **Uncontaminated.** Candidates are drawn from the upstream sources (Aya: `train-0.jsonl` and `test-0.jsonl`[3]; WildChat: `train-00.jsonl` through `train-13.jsonl`), subject to the 13-gram filter, with the token budget allocated proportionally to the two subsets. Because the 1B/32B and 7B/13B models use distinct SFT training mixes, uncontaminated splits are constructed separately for each pair. **Contaminated.** For each model pair, contaminated candidates are drawn from the corresponding OLMo 2 SFT training data, restricted to instances originally from WildChat or Aya.

Post-training/DPO. We use the WildChat subset alone (see rationale in Appendix G). **Uncontaminated.** Preference data are constructed in three steps: (i) we sample 400 prompts from original WildChat (`train-00.jsonl` to `train-13.jsonl`) that pass the 13-gram filter; (ii) we generate chosen and rejected responses using the UltraFeedback pipeline (Cui et al., 2024) following OLMo 2; (iii) we retain only instances where the prompt, chosen, and rejected responses all pass the 13-gram filter, as model-generated outputs may overlap with OLMo 2's training data. From the resulting uncontaminated pool, we sample approximately 217K tokens. Because the 1B, 7B, 13B, and 32B models use different DPO mixes, uncontaminated splits are constructed separately for each. **Contaminated.** For each model size, candidates are drawn from the corresponding OLMo 2 DPO training data, restricted to instances originating from WildChat.

Post-training/RLVR. RLVR consists of two subsets: math-verifier (GSM8K with 8-shot CoT and MATH with 3-shot CoT) and IFEval, with a target token ratio of approximately 2.3:1 (math-verifier to IFEval). **Uncontaminated.** For math-verifier, candidates are drawn from `gsm8k-test.jsonl` (GSM8K) and the test split within `train-00.jsonl` (MATH; train/test

[3]We treat both files together as the upstream Aya corpus, since OLMo 2's Aya source is reported only at the aggregate level (202,285 prompts) without per-file attribution.

instances are separated during preprocessing), and retained if they pass the 13-gram filter, then prepended with the same CoT exemplars used in OLMo 2's RLVR setup. For IFEval, prompts are drawn from the Tülu 2 SFT Mixture (Ivison et al., 2023), filtered using the 13-gram criterion, and augmented with the same verifiable instruction-following templates used in OLMo 2's RLVR setup. **Contaminated.** Contaminated math-verifier and IFEval instances are drawn separately from OLMo 2's RLVR training data (`train-00.jsonl`).

F Source-Matched Candidate Pools for Post-training Domains

This appendix details the source-matched candidate pools for the three post-training domains. Unlike DCLM-Baseline, StarCoder, and Stack Exchange, where OLMo 2 applies domain-specific quality filters, the three post-training domains are drawn directly from their upstream source corpora without further quality filtering.

Post-training/SFT. Six subsets of OLMo 2's SFT mix are randomly downsampled from larger upstream corpora: OpenAssistant, WildChat, SciRIFF, TableGPT, OpenMathInstruct 2, and Aya. See Table 12 for per-subset prompt counts.

Post-training/DPO. Three of OLMo 2's DPO prompt subsets are randomly sampled from larger corpora: WildChat, DaringAnteater, and Tülu 3 Personas IF (see Table 13). Source-matched preference data can be obtained by sampling uncontaminated prompts from their remainders and applying OLMo 2's UltraFeedback pipeline for response generation and preference annotation.

Post-training/RLVR (IFEval). OLMo 2's RLVR IFEval prompts are randomly sampled from the Tülu 2 SFT mix (Ivison et al., 2023) and augmented with verifiable instruction-following constraints. Source-matched uncontaminated candidates can be obtained by sampling from the remainder of Tülu 2 SFT mix and applying the same constraint procedure.

G Per-Model Training Corpora for Infini-Gram Indexing

We construct 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.

Notably, OLMo 2's pre-training uses non-integer epoch counts (Appendix C.1): 1 epoch for 1B and 7B, 1.2 epochs for 13B, and 1.5 epochs for 32B. Our index reflects only the integer-epoch portion, introducing a worst-case error of ± 1 in n-gram occurrence counts.

H Subset Selection for Multi-Subset Domains

This appendix details the subset-selection strategy for each domain spanning multiple sampleable subsets. We select two subsets per domain with budgets proportional to their training-token counts. Moreover, to ensure sufficient evaluation signal from each subset, we cap the allocation ratio at 1:20, as more extreme ratios would leave the smaller subset underrepresented.

Pre-training/StarCoder. OLMo 2's filtered StarCoder corpus spans 86 programming languages. We select Java and Assembly, representing a high-resource general-purpose language and a low-resource systems language respectively.

Mid-training/Stack Exchange. OLMo 2's Stack Exchange corpus comprises Q&A pairs from 362 forums. We select Stack Overflow (the largest forum, primarily programming Q&A) and Math (a mid-sized math-focused forum).

Mid-training/Dolmino Math Mix. Although Dolmino Math Mix contains multiple subsets, only the GSM8K subset has held-out data available; the remaining subsets lack both held-out splits and larger source corpora. Therefore, we treat GSM8K as the only sampleable subset for this domain.

Post-training/SFT. Six of OLMo 2's SFT subsets are downsampled from larger upstream corpora and have sampleable remainders: Aya, WildChat, OpenAssistant, SciRIFF, TableGPT, and OpenMathInstruct 2 (see Table 12 for details). We select WildChat (Zhao et al., 2024), the largest by training tokens and a corpus of real-user conversations with ChatGPT, and Aya (Singh et al., 2024), the second-largest multilingual instruction-following dataset spanning over 100 languages.

Post-training/DPO. Three of OLMo 2's DPO prompt subsets are sampleable: WildChat, DaringAnteater, and Tülu 3 Personas IF (see Table 13). We select WildChat as the primary sub-

set due to its dominant training footprint (84,105 prompts). DaringAnteater is excluded, as its 1,618 prompts in OLMo 2's DPO mix induce a training-token ratio of approximately 1:52 with WildChat, exceeding the 1:20 cap. Under proportional allocation, this would assign fewer than 10 evaluation prompts to DaringAnteater, which is insufficient for stable AUC estimation. Increasing its allocation by sampling additional uncontaminated data from its source remainder would over-represent it relative to its original training mixture, violating source-matched proportionality. Tülu 3 Personas IF was also attempted but yields fewer than 10 uncontaminated prompts after 13-gram filtering. As a result, the DPO domain retains WildChat alone.

Post-training/RLVR. RLVR comprises three subsets: 8-shot CoT-prepended GSM8K, 3-shot CoT-prepended MATH, and IFEval. We retain all three rather than subsampling, as they follow fundamentally different task structures.

I Constrained Sampling via Simulated Annealing

This appendix provides a detailed description of the simulated-annealing procedure used to construct contaminated splits in Section 3.5.

Setup. Given a candidate pool P and a reference uncontaminated split with token distribution q, the goal is to select a subset $D \subseteq P$ satisfying three constraints: (i) $|D|$ equals the uncontaminated instance count; (ii) the total token count of D is within ± 250 of the uncontaminated token count; and (iii) the distributional loss $\mathcal{L}$ against the uncontaminated reference is minimized:

$$\mathcal{L} = \mathrm{JSD}(p, q) + \mathbb{1}[\text{temporal metadata}] \cdot \mathcal{L}_{\text{temp}}, \tag{1}$$

where p is the token distribution of D, JSD denotes the Jensen-Shannon divergence enforcing lexical alignment, and the temporal alignment term takes the form:

$$\mathcal{L}_{\text{temp}} = \frac{|\mu_p - \mu_q| + |\sigma_p - \sigma_q|}{\sigma_q}, \tag{2}$$

with μ and σ denoting the mean and standard deviation of the temporal axis (in days), normalized by the uncontaminated standard deviation σ_q.

Algorithm. We initialize by randomly partitioning P into $(D, \text{leftover})$ such that $|D|$ matches the target instance count. The algorithm then runs for $T_{\max} = 10^7$ steps, with each step proposing a swap between one instance in D 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 ± 250 margin, the swap is accepted only if it brings the token count closer to target. In the **alignment optimization phase**, used once in margin, the swap is accepted if the post-swap token count remains in 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 42, and the resulting contaminated split is retained as the final output.

J Detailed 7-Gram Overlap between Uncontaminated and Contaminated Splits

As illustrated in Table 23, 7-gram overlap remains uniformly low across all four OLMo 2 model sizes (1B, 7B, 13B, and 32B), with the maximum overlap reaching just 0.06502 (GSM8K). Pre-training, mid-training, and RLVR values are identical across columns since their splits are shared, while SFT and DPO values vary slightly due to model-specific training mixes.

K Detailed Per-Domain Statistics of OLMo-Detect (Shifted)

We provide detailed statistics for OLMo-Detect (Shifted). Per-domain statistics for the pre-training, mid-training, and post-training stages are reported in Tables 24–26; per-subset statistics for StarCoder, Stack Exchange, and SFT are reported in Tables 27–29.

Across the seven included domains, contaminated splits exhibit substantially larger mismatches in text quality and temporal range than those in OLMo-Detect; lexical (JSD) mismatches also increase, though the magnitude varies by domain (e.g., Stack Exchange: 0.05443 → 0.06374 vs. OpenWebMath: 0.07519 → 0.22611). GSM8K and RLVR are excluded, since random sampling

does not yield distributionally shifted contaminated splits in these domains.

Token Count Margin. OLMO-DETECT (SHIFTED) matches token counts within ± 500 of the uncontaminated counterparts, slightly wider than the ± 250 margin used in OLMO-DETECT. This relaxation is necessary because random sampling has higher token-count variance than constrained sampling via simulated annealing. The resulting deviation remains under 0.25% of the 217K-token budget per domain.

Dev/Test Variance. Unlike in OLMO-DETECT, the dev and test partitions of contaminated splits in OLMO-DETECT (SHIFTED) can exhibit visible discrepancies (e.g., DCLM-Baseline FastText: 0.13630 in dev vs. 0.16893 in test). In OLMO-DETECT, simulated annealing optimizes the contaminated split as a whole to match the uncontaminated reference, yielding similar distributions across its dev and test sub-partitions. However, OLMO-DETECT (SHIFTED) applies no such optimization: contaminated samples are drawn uniformly at random from OLMo 2's training data, which exhibits substantial variance along quality and temporal axes (e.g., DCLM-Baseline FastText scores have $\sigma \approx 0.24$ in the contaminated split). Combined with small dev set sizes (15–44 in pretraining and mid-training), random partitioning can produce visible dev–test drift due solely to sampling variance. This behavior is expected by design and does not invalidate the evaluation: dev and test are drawn from the same underlying distribution; consequently, the ability of methods to generalize from dev to test under distributional shift remains meaningfully measurable.

L Detailed Experimental Settings

Excluded Baselines. Among existing detection methods, seven are excluded due to setting incompatibility: Smaller Ref (Carlini et al., 2021) and SPV-MIA (Fu et al., 2024) require a separately trained reference model; Permutation Test (Oren et al., 2023) and Local Order Quiz (Samuel et al., 2025) require canonical dataset ordering, which is incompatible with OLMo 2's randomly shuffled training; Name-cloze (Chang et al., 2023) requires text containing proper nouns to mask; TS-Guessing (Deng et al., 2024) has no open-source implementation; ConStat (Dekoninck et al., 2024) requires a comparable-difficulty reference benchmark.

Implementation Details. All compared methods follow their original GitHub implementations. To ensure fair comparison, all experiments are conducted on two A100 GPUs with a context length of 4096. Hyperparameters, where applicable, are tuned on the dev set. All experiments use a fixed random seed 42.

Continuous Scoring for Categorical Methods. DCQ (Golchin and Surdeanu, 2025) and Guided Instruction (Golchin and 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 $P(L \mid \mathrm{BCQ}_L)$ 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 Instruction, 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 Instruction); only the post-processing step that maps these signals to a categorical decision is replaced with continuous score extraction.

M Per-Domain AUC Results on OLMO-DETECT

This appendix provides detailed AUC results on OLMO-DETECT for each of the nine domains (Section 3.3), complementing the stage-level aggregates in Table 4. Per-domain AUC results are shown in Tables 30–42. Each table reports per-method AUC across the four OLMo 2 model sizes, with the Model-Avg row averaging across sizes and the Method-Avg column averaging across methods.

Degenerate scores on CoT-prefixed RLVR subsets. On the RLVR/CoT-GSM8K subset, Neighborhood Attack (Neigh.) and EMMIA both return an AUC of exactly 0.5000 across all four model sizes (Table 42); EMMIA shows the same pattern on RLVR/CoT-MATH. Such exactness across sizes indicates degenerate output rather than statistical chance: the underlying scoring procedure produces no discriminative signal and defaults to a constant. The failure tracks sample length: both methods fail on the 8-shot CoT-GSM8K subset, EMMIA additionally fails on the 3-shot CoT-MATH subset, while both produce non-degenerate scores on the prefix-free IFEval subset. We attribute this to limitations in the methods' scoring procedures on long, CoT-prefixed samples: Neighborhood Attack's perturbations may disrupt the Question/Reasoning/Answer template rather than the unique content, eliminating the loss-difference signal it relies on; EMMIA's iterative scoring procedure may exceed its calibrated input range. We retain these values verbatim for transparency but do not interpret them as meaningful detection performance. The same pattern is observed in TPR@5%FPR (Table 59): both methods default to exactly 0.0500 (the 5%FPR chance level) on the same subsets.

N TPR@5%FPR Results on OLMo-Detect

This appendix provides detailed TPR@5%FPR results on OLMo-Detect, following the same structure as Appendix M. Aggregated, stage-wise and per-domain results are shown in Tables 43–59. Overall trends largely mirror those observed under AUC: detection is strongest at mid-training, moderate at pre-training, and weakest at post-training (Method-Avg of 0.27, 0.12, 0.10 respectively); CAMIA remains the strongest method overall; and performance improves with model size before plateauing. Two differences are worth noting: (i) TPR@5%FPR is more sensitive to memorization strength than AUC, so cross-stage gaps appear sharper (mid-training is roughly 2× pre-training under TPR but only 1.05× under AUC); and (ii) the size-scaling trend is closer to monotonic under TPR (1B/7B/13B/32B Method-Avg of 0.10/0.15/0.22/0.23) than under AUC, suggesting that larger models surface more high-confidence memorization signal even when the rank-ordering metric saturates.

O Domain-Specific Structural Framing: Illustrations and Caveats

This appendix complements Section 5.2 by (i) illustrating the structure of domains *without* domain-specific structural framing, and (ii) explaining the omission of StarCoder/Assembly from Table 6.

Free-form domains and apparent-but-shared framing. Figure 1 (main text) shows two domains with domain-specific structural framing. The remaining seven domains in OLMo-Detect either lack any recurring per-sample format (*free-form*) or carry recurring tokens that are not domain-specific. Figure 2 shows representative samples from each.

As shown in Figure 2, the pre-training domains DCLM-Baseline, peS2o, and OpenWebMath, as well as the mid-training domain Stack Exchange, are free-form: every sample is a self-contained passage with no domain-wide scaffold (panels (a)–(c)). The three post-training domains SFT, DPO, and RLVR (panel (d)) do carry recurring tokens, namely the chat-template delimiters `<|user|>` and `<|assistant|>`, but these tokens appear *identically* across all three domains, so they cannot serve to anchor any particular domain's content. By the definition in Section 5.2, framing must be unique to the domain to count as domain-specific.

Omission of StarCoder/Assembly. The pre-training domain StarCoder/Assembly is omitted from Table 6 due to its small sample size. As shown in Table 18, StarCoder/Java contains 344 contaminated samples (309 evaluation + 35 development), whereas StarCoder/Assembly contains only 21 (18 evaluation + 3 development), an order of magnitude fewer. This disparity reflects our proportional sampling relative to each subset's contribution to the OLMo 2 training corpus (Appendix H). StarCoder/Assembly is also a borderline case for structural framing: while it lacks Java's universal `package`/`import`/`public class` templates, its samples (largely OEIS-derived) share weaker recurring patterns such as `; A####:` comment markers and comma-separated number sequences. Combined with the scarce samples, the relative contribution of these two factors to its observed Model-Avg AUC of 0.5453 (Table 39), well below Java's 0.7521, cannot be reliably isolated.

(a) Pre-training/DCLM-Baseline, Avg AUC 0.56

```
The INS and Illegal Aliens: an Old, Distorted Story.
October 19, 1986. SAN DIEGO – Congressional agreement
last week on landmark immigration legislation ...
```

(b) Pre-training/peS2o, Avg AUC 0.53

```
Microbial Processes in Engineering Clay Materials
and Biocidal Additives to Prevent Them. Microbial
activity in engineering clay materials of safety ...
```

(c) Mid-training/Stack Exchange, Avg AUC 0.61

How many numbers of 4 digits $x_1x_2x_3x_4$ satisfy $x_1 \leq x_2 \leq x_3 \leq x_4$? I have identified this situation with distributing 4 equal objects ...

(d) Post-training/SFT, DPO, RLVR, Avg AUC 0.59/0.50/0.56

```
<|user|> What is the answer to ...?
<|assistant|> The answer is ...
```

Figure 2: Representative samples from domains *without* domain-specific structural framing. Panels (a)–(c) are free-form: unique content varies per sample (yellow), with no recurring scaffold. Panel (d) shows the chat-template delimiters shared across SFT, DPO, and RLVR (gray): although the delimiters recur within each sample, they are identical across all three domains and therefore not domain-specific.

P Per-Stage and Per-Domain Matched-vs-Shifted Results

This appendix provides detailed matched-vs-shifted AUC comparisons on OLMo-Detect and OLMo-Detect (Shifted), complementing the aggregate results in Table 8. Results are reported in Tables 60–72. Each table follows the same layout as Table 8: four model sizes (1B, 7B, 13B, 32B) under both *Matched* and *Shifted* settings, with per-method **Avg** Δ, p-value, and **Class.** (R/I/D) summary rows. GSM8K (mid-training) and RLVR (post-training) are omitted from per-domain tables, as no distribution-shifted counterpart exists for these domains. SelfCrit's results on DPO are shown in Table 69.

Trends largely mirror the aggregate. The three findings in Section 5.5 hold consistently across training stages: (i) MinK and DCPDD inflate across all stages, with shifted AUC approaching 1.00 (e.g., DCPDD reaches 0.9979–1.0000 on shifted across all stages and sizes); (ii) Lower degrades consistently in pre-training and remains robust elsewhere; Zlib shows a similar pattern except on StarCoder (Table 66), where the compressed-length normalization remains stable under code-vs-natural-text shift, leading to inflation rather than degradation; (iii) Method-Avg shifts upward overall (Avg Δ ranging from $+0.05$ to $+0.10$ per stage), classified as inflated.

Stage-dependent classifications. A few methods change classification across stages. Neighborhood Attack (Neigh.) is *degraded* on pre-training (Avg $\Delta = -0.08$) but *robust* on post-training; conversely, RECALL and CAMIA shift from robust on pre-training to inflated on mid-training and post-training. These stage-level differences reflect domain-composition effects: each stage contains a different mix of free-form and structurally-framed domains, and the shifted data differs accordingly.

Surprising patterns. We note two patterns worth flagging. First, on the StarCoder/Assembly subset (Table 70), CAMIA produces an exact 0.0000 AUC on the shifted setting at the 7B size, signaling a complete score inversion (non-members consistently scoring higher than members), likely due to the small sample size ($N = 21$, as discussed in Appendix O). Second, DCPDD's shifted AUC saturates at exactly 1.0000 on multiple stage and domain combinations, indicating perfect separability driven by the token-probability discrepancy between in-distribution members and out-of-distribution non-members, a clear instance of detection performance reflecting topic discrimination rather than memorization.

Interpretation of robustness classifications. The **R** classification reflects $p > 0.10$ under a paired t-test with df $= 3$ and should be interpreted as "no detected shift" rather than proof of true invariance. With only four model sizes, the test has limited power to detect small distributional shifts; borderline cases (small $|\overline{\Delta}|$ with p near 0.10) should be interpreted with appropriate caution.

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 uncontaminated candidates available for OLMO-DETECT: ✓ = sampleable remainder available; blank = source corpus fully consumed (# in SFT = Count).

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.

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).

Stage	Domain	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline	Uncon	Dev	16	21,745	FastText: 0.01808 ± 0.00004	2018-01-18 ± 930d	0.05090
			Test	134	195,372	FastText: 0.01809 ± 0.00008	2018-01-19 ± 930d	
			Total	**150**	**217,117**			
		Con	Dev	16	21,721	FastText: 0.01811 ± 0.0	2018-01-17 ± 932d	
			Test	134	195,369	FastText: 0.01811 ± 0.0	2018-01-17 ± 932d	
			Total	**150**	**217,090**			
	peS2o	Uncon	Dev	27	21,746	—	2022-12-15 ± 12d	0.07321
			Test	241	195,465	—	2022-12-15 ± 14d	
			Total	**268**	**217,211**			
		Con	Dev	27	21,804	—	2022-12-15 ± 13d	
			Test	241	195,426	—	2022-12-15 ± 13d	
			Total	**268**	**217,230**			
	OpenWebMath	Uncon	Dev	15	21,761	—	2019-10-12 ± 850d	0.07519
			Test	136	195,422	—	2019-10-12 ± 852d	
			Total	**151**	**217,183**			
		Con	Dev	15	21,793	—	2019-10-14 ± 851d	
			Test	136	195,571	—	2019-10-17 ± 863d	
			Total	**151**	**217,364**			
	StarCoder	Uncon	Dev	38	21,770	GitHub Stars: 159.45 ± 523.86 Top-1 Freq: 0.30984 ± 0.00913 Top-2 Freq: 0.08653 ± 0.03881	—	0.17624
			Test	327	195,572	GitHub Stars: 159.75 ± 524.62 Top-1 Freq: 0.31010 ± 0.00837 Top-2 Freq: 0.07911 ± 0.03699	—	
			Total	**365**	**217,342**			
		Con	Dev	38	21,701	GitHub Stars: 159.50 ± 524.01 Top-1 Freq: 0.29199 ± 0.00602 Top-2 Freq: 0.06633 ± 0.04123	—	
			Test	327	195,566	GitHub Stars: 160.07 ± 524.53 Top-1 Freq: 0.28910 ± 0.00600 Top-2 Freq: 0.05449 ± 0.02036	—	
			Total	**365**	**217,267**			

Table 15: Detailed per-domain statistics for the pre-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	GSM8K	Uncon	Dev	103	21,755	—	—	0.02585
			Test	937	195,408	—	—	
			Total	**1,040**	**217,163**			
		Con	Dev	103	21,729	—	—	
			Test	937	195,403	—	—	
			Total	**1,040**	**217,132**			
	Stack Exchange	Uncon	Dev	44	21,744	Question Vote: 3.43 ± 2.10 Answer Vote: 5.11 ± 2.06	2014-12-04 ± 1325d	0.05443
			Test	395	195,397	Question Vote: 3.43 ± 2.09 Answer Vote: 5.11 ± 2.04	2014-12-04 ± 1326d	
			Total	**439**	**217,141**			
		Con	Dev	44	21,849	Question Vote: 3.75 ± 1.80 Answer Vote: 6.00 ± 2.17	2014-12-03 ± 1324d	
			Test	395	195,303	Question Vote: 4.08 ± 2.10 Answer Vote: 6.21 ± 2.06	2014-12-08 ± 1303d	
			Total	**439**	**217,152**			

Table 16: Sampling statistics for the mid-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Model	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Uncon	Dev	21	21,697	—	—	0.08951
				Test	204	195,546	—	—	
				Total	**225**	**217,243**			
			Con	Dev	21	21,696	—	—	
				Test	204	195,317	—	—	
				Total	**225**	**217,013**			
		7B, 13B	Uncon	Dev	20	21,640	—	—	0.08338
				Test	148	195,237	—	—	
				Total	**168**	**216,877**			
			Con	Dev	20	21,956	—	—	
				Test	148	195,420	—	—	
				Total	**168**	**217,376**			
	DPO	1B	Uncon	Dev	11	21,723	—	—	0.05755
				Test	102	195,418	—	—	
				Total	**113**	**217,141**			
			Con	Dev	11	21,722	—	—	
				Test	102	195,565	—	—	
				Total	**113**	**217,287**			
		7B	Uncon	Dev	10	21,726	—	—	0.09247
				Test	95	195,473	—	—	
				Total	**105**	**217,199**			
			Con	Dev	10	21,725	—	—	
				Test	95	195,590	—	—	
				Total	**105**	**217,315**			
		13B	Uncon	Dev	11	21,695	—	—	0.08240
				Test	99	195,498	—	—	
				Total	**110**	**217,193**			
			Con	Dev	11	21,694	—	—	
				Test	99	195,565	—	—	
				Total	**110**	**217,259**			
		32B	Uncon	Dev	10	21,822	—	—	0.08469
				Test	94	195,393	—	—	
				Total	**104**	**217,215**			
			Con	Dev	10	21,935	—	—	
				Test	94	195,394	—	—	
				Total	**104**	**217,329**			
	RLVR	1B, 7B, 13B, 32B	Uncon	Dev	41	21,745	—	—	0.04473
				Test	369	195,420	—	—	
				Total	**410**	**217,165**			
			Con	Dev	41	21,577	—	—	
				Test	369	195,415	—	—	
				Total	**410**	**216,992**			

Table 17: Detailed per-stage statistics for the post-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	StarCoder	Assembly	Uncon	Dev	3	1,098	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.30705 ± 0.00269 Top-2 Freq: 0.05158 ± 0.01214	—	0.01720
				Test	18	9,837	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.30583 ± 0.00269 Top-2 Freq: 0.05030 ± 0.01345	—	
				Total	**21**	**10,935**			
			Con	Dev	3	1,114	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.29924 ± 0.00021 Top-2 Freq: 0.05162 ± 0.01249	—	
				Test	18	9,824	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.29929 ± 0.00017 Top-2 Freq: 0.07691 ± 0.02013	—	
				Total	**21**	**10,938**			
		Java	Uncon	Dev	35	20,672	GitHub Stars: 171.86 ± 544.06 Top-1 Freq: 0.31008 ± 0.00944 Top-2 Freq: 0.08953 ± 0.03885	—	0.19607
				Test	309	185,735	GitHub Stars: 168.20 ± 538.48 Top-1 Freq: 0.31035 ± 0.00852 Top-2 Freq: 0.08079 ± 0.03724	—	
				Total	**344**	**206,407**			
			Con	Dev	35	20,587	GitHub Stars: 171.91 ± 544.21 Top-1 Freq: 0.29136 ± 0.00587 Top-2 Freq: 0.06759 ± 0.04256	—	
				Test	309	185,742	GitHub Stars: 168.54 ± 538.38 Top-1 Freq: 0.28851 ± 0.00563 Top-2 Freq: 0.05318 ± 0.01960	—	
				Total	**344**	**206,329**			

Table 18: Detailed per-subset statistics for StarCoder in the pre-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Math Stack Exchange	Uncon	Dev	4	2,011	Question Vote: 3.25 ± 1.64 Answer Vote: 4.75 ± 0.83	2015-01-02 ± 960.79d	0.11183
				Test	35	18,128	Question Vote: 3.23 ± 1.67 Answer Vote: 4.71 ± 0.85	2015-01-01 ± 977.82d	
				Total	**39**	**20,139**			
			Con	Dev	4	2,026	Question Vote: 3.75 ± 1.30 Answer Vote: 5.25 ± 0.43	2015-01-04 ± 961.37d	
				Test	35	18,032	Question Vote: 4.00 ± 1.97 Answer Vote: 5.34 ± 0.75	2014-12-30 ± 976.38d	
				Total	**39**	**20,058**			
		Stack Overflow	Uncon	Dev	40	19,733	Question Vote: 3.45 ± 2.14 Answer Vote: 5.15 ± 2.14	2014-12-01 ± 1356.23d	0.05719
				Test	360	177,269	Question Vote: 3.45 ± 2.13 Answer Vote: 5.15 ± 2.11	2014-12-01 ± 1355.68d	
				Total	**400**	**197,002**			
			Con	Dev	40	19,823	Question Vote: 3.75 ± 1.84 Answer Vote: 6.08 ± 2.26	2014-11-30 ± 1354.67d	
				Test	360	177,271	Question Vote: 4.08 ± 2.11 Answer Vote: 6.29 ± 2.13	2014-12-05 ± 1330.24d	
				Total	**400**	**197,094**			

Table 19: Detailed per-subset statistics for Stack Exchange in the mid-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Aya	Uncon	Dev	14	5,293	—	—	0.05355
					Test	140	47,321	—	—	
					Total	**154**	**52,614**			
				Con	Dev	14	5,258	—	—	
					Test	140	47,199	—	—	
					Total	**154**	**52,457**			
			Wildchat	Uncon	Dev	7	16,404	—	—	0.11885
					Test	64	148,225	—	—	
					Total	**71**	**164,629**			
				Con	Dev	7	16,438	—	—	
					Test	64	148,118	—	—	
					Total	**71**	**164,556**			
		7B, 13B	Aya	Uncon	Dev	13	5,117	—	—	0.07653
					Test	82	46,860	—	—	
					Total	**95**	**51,977**			
				Con	Dev	13	5,301	—	—	
					Test	82	47,059	—	—	
					Total	**95**	**52,360**			
			Wildchat	Uncon	Dev	7	16,523	—	—	0.10279
					Test	66	148,377	—	—	
					Total	**73**	**164,900**			
				Con	Dev	7	16,655	—	—	
					Test	66	148,361	—	—	
					Total	**73**	**165,016**			

Table 20: Detailed per-subset statistics for SFT in the post-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	RLVR	1B, 7B, 13B, 32B	CoT-GSM8K	Uncon	Dev	9	7,611	—	—	0.01128
					Test	88	73,906	—	—	
					Total	**97**	**81,517**			
				Con	Dev	9	7,541	—	—	
					Test	88	73,971	—	—	
					Total	**97**	**81,512**			
			CoT-MATH	Uncon	Dev	11	7,525	—	—	0.01497
					Test	93	61,897	—	—	
					Total	**104**	**69,422**			
				Con	Dev	11	7,428	—	—	
					Test	93	61,793	—	—	
					Total	**104**	**69,221**			
			IFEval	Uncon	Dev	21	6,609	—	—	0.14588
					Test	188	59,617	—	—	
					Total	**209**	**66,226**			
				Con	Dev	21	6,608	—	—	
					Test	188	59,651	—	—	
					Total	**209**	**66,259**			

Table 21: Detailed per-subset statistics for RLVR in the post-training stage of OLMO-DETECT. **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Uncontaminated		Contaminated		
		Held-Out	Remainder of Original	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline		✓	✓	✓	✓
	peS2o	✓			✓	✓
	OpenWebMath	✓			✓	✓
	StarCoder		✓	✓		✓
Mid-training	Stack Exchange		✓	✓	✓	✓
	Dolmino Math Mix†	✓				✓
Post-training	SFT		✓			✓
	DPO		✓			✓
	RLVR‡	✓	✓			✓

Table 22: Construction methodology for OLMO-DETECT. **Uncontaminated** samples are sampled from either **Held-Out** data or the **Remainder of Original** source corpus from which the contaminated counterpart was drawn, followed by rigorous filtering against the OLMo 2 training corpus via infini-gram. **Contaminated** samples are sampled from OLMo 2 training data to explicitly match their uncontaminated counterparts along up to three axes: Text **Quality**, **Temporal** Range, and **Lexical** Similarity. †For Dolmino Math Mix, sampling is applied specifically to the GSM8K subset. ‡For RLVR, held-out sampling is applied to CoT-GSM8K and CoT-MATH subsets, while remainder-of-original sampling is applied to the IFEval subset.

Stage	Domain	1b	7b	13b	32b	Avg
Pre-training	DCLM-Baseline	0.00117	0.00117	0.00117	0.00117	0.00117
	OpenWebMath	0.00242	0.00242	0.00242	0.00242	0.00242
	peS2o	0.00193	0.00193	0.00193	0.00193	0.00193
	StarCoder	0.01574	0.01574	0.01574	0.01574	0.01574
Mid-training	GSM8K	0.06502	0.06502	0.06502	0.06502	0.06502
	Stack Exchange	0.00402	0.00402	0.00402	0.00402	0.00402
Post-training	SFT	0.01077	0.01374	0.01374	0.01077	0.01226
	DPO	0.01044	0.01049	0.01168	0.00940	0.01050
	RLVR	0.03109	0.03109	0.03109	0.03109	0.03109

Table 23: Per-model 7-gram overlap between contaminated ($\mathcal{C}$) and uncontaminated ($\mathcal{U}$) splits in OLMO-DETECT, computed as $|\mathcal{C} \cap \mathcal{U}|/|\mathcal{C}|$. The **Avg** column reports the arithmetic mean across the 1B, 7B, 13B, and 32B models. Pre-training, mid-training, and RLVR splits are shared across model sizes, yielding identical values across columns. SFT splits are shared within the 1B/32B and 7B/13B model pairs, whereas DPO splits are distinct for each size.

Stage	Domain	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	DCLM-Baseline	Uncon	Dev	16	21,745	FastText: 0.01808 ± 0.00004	2018-01-18 ± 930d	0.08300
			Test	134	195,372	FastText: 0.01809 ± 0.00008	2018-01-19 ± 930d	
			Total	**150**	**217,117**			
		Con	Dev	16	21,690	FastText: 0.13630 ± 0.23571	2019-01-23 ± 0d	
			Test	134	195,388	FastText: 0.16893 ± 0.23648	2019-01-22 ± 0d	
			Total	**150**	**217,078**			
	peS2o	Uncon	Dev	27	21,746	—	2022-12-15 ± 12d	0.09198
			Test	241	195,465	—	2022-12-15 ± 14d	
			Total	**268**	**217,211**			
		Con	Dev	27	21,993	—	2014-02-03 ± 2656d	
			Test	241	195,630	—	2014-02-02 ± 2657d	
			Total	**268**	**217,623**			
	OpenWebMath	Uncon	Dev	15	21,761	—	2019-10-12 ± 850d	0.22611
			Test	136	195,422	—	2019-10-12 ± 852d	
			Total	**151**	**217,183**			
		Con	Dev	15	21,820	—	2011-10-01 ± 2943d	
			Test	136	195,571	—	2011-09-28 ± 2940d	
			Total	**151**	**217,391**			
	StarCoder	Uncon	Dev	38	21,770	GitHub Stars: 159.45 ± 523.86 Top-1 Freq: 0.30984 ± 0.00913 Top-2 Freq: 0.08653 ± 0.03881	—	0.27694
			Test	327	195,451	GitHub Stars: 159.68 ± 524.64 Top-1 Freq: 0.31005 ± 0.00837 Top-2 Freq: 0.07930 ± 0.03701	—	
			Total	**365**	**217,221**			
		Con	Dev	38	21,850	GitHub Stars: 345.58 ± 1767.44 Top-1 Freq: 0.08330 ± 0.03806 Top-2 Freq: 0.05711 ± 0.01598	—	
			Test	327	195,834	GitHub Stars: 350.17 ± 1790.63 Top-1 Freq: 0.08162 ± 0.03459 Top-2 Freq: 0.05681 ± 0.01562	—	
			Total	**365**	**217,684**			

Table 24: Detailed per-domain statistics for the pre-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Uncon	Dev	44	21,744	Question Vote: 3.43 ± 2.10 Answer Vote: 5.11 ± 2.06	2014-12-04 ± 1325d	0.06374
			Test	395	195,397	Question Vote: 3.43 ± 2.09 Answer Vote: 5.11 ± 2.04	2014-12-04 ± 1326d	
			Total	**439**	**217,141**			
		Con	Dev	44	21,900	Question Vote: 17.52 ± 40.18 Answer Vote: 25.20 ± 56.55	2013-06-04 ± 1069d	
			Test	395	195,517	Question Vote: 19.19 ± 40.33 Answer Vote: 26.76 ± 57.45	2013-05-25 ± 1069d	
			Total	**439**	**217,417**			

Table 25: Detailed per-domain statistics for the mid-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Model	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Uncon	Dev	21	21,697	—	—	0.21675
				Test	204	195,546	—	—	
				Total	**225**	**217,243**			
			Con	Dev	21	21,298	—	—	
				Test	204	195,511	—	—	
				Total	**225**	**216,809**			
		7B, 13B	Uncon	Dev	20	21,640	—	—	0.27042
				Test	148	195,237	—	—	
				Total	**168**	**216,877**			
			Con	Dev	20	21,716	—	—	
				Test	148	195,602	—	—	
				Total	**168**	**217,318**			
	DPO	1B	Uncon	Dev	11	21,723	—	—	0.21413
				Test	102	195,418	—	—	
				Total	**113**	**217,141**			
			Con	Dev	11	21,876	—	—	
				Test	102	195,600	—	—	
				Total	**113**	**217,476**			
		7B	Uncon	Dev	10	21,726	—	—	0.21567
				Test	95	195,473	—	—	
				Total	**105**	**217,199**			
			Con	Dev	10	21,725	—	—	
				Test	95	195,366	—	—	
				Total	**105**	**217,091**			
		13B	Uncon	Dev	11	21,695	—	—	0.21412
				Test	99	195,498	—	—	
				Total	**110**	**217,193**			
			Con	Dev	11	21,456	—	—	
				Test	99	195,870	—	—	
				Total	**110**	**217,326**			
		32B	Uncon	Dev	10	21,822	—	—	0.21440
				Test	94	195,393	—	—	
				Total	**104**	**217,215**			
			Con	Dev	10	21,778	—	—	
				Test	94	195,758	—	—	
				Total	**104**	**217,536**			

Table 26: Detailed per-stage statistics for the post-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Pre-training	StarCoder	Assembly	Uncon	Dev	3	1,098	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.30705 ± 0.00269 Top-2 Freq: 0.05158 ± 0.01214	—	0.24219
				Test	18	9,837	GitHub Stars: 14.67 ± 5.19 Top-1 Freq: 0.30583 ± 0.00269 Top-2 Freq: 0.05030 ± 0.01345	—	
				Total	**21**	**10,935**			
			Con	Dev	3	1,124	GitHub Stars: 178.00 ± 242.57 Top-1 Freq: 0.14517 ± 0.08277 Top-2 Freq: 0.06724 ± 0.02305	—	
				Test	18	10,035	GitHub Stars: 180.50 ± 260.10 Top-1 Freq: 0.14331 ± 0.08318 Top-2 Freq: 0.06717 ± 0.02292	—	
				Total	**21**	**11,159**			
		Java	Uncon	Dev	35	20,672	GitHub Stars: 171.86 ± 544.06 Top-1 Freq: 0.31008 ± 0.00944 Top-2 Freq: 0.08953 ± 0.03885	—	0.29834
				Test	309	185,614	GitHub Stars: 168.20 ± 538.48 Top-1 Freq: 0.31035 ± 0.00852 Top-2 Freq: 0.08079 ± 0.03724	—	
				Total	**344**	**206,286**			
			Con	Dev	35	20,726	GitHub Stars: 359.94 ± 1839.55 Top-1 Freq: 0.07800 ± 0.02508 Top-2 Freq: 0.05624 ± 0.01490	—	
				Test	309	185,799	GitHub Stars: 360.06 ± 1840.49 Top-1 Freq: 0.07803 ± 0.02506 Top-2 Freq: 0.05621 ± 0.01486	—	
				Total	**344**	**206,525**			

Table 27: Detailed per-subset statistics for StarCoder in the pre-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Mid-training	Stack Exchange	Math Stack Exchange	Uncon	Dev	4	2,011	Question Vote: 3.25 ± 1.64 Answer Vote: 4.75 ± 0.83	2015-01-02 ± 960.79d	0.12285
				Test	35	18,128	Question Vote: 3.23 ± 1.67 Answer Vote: 4.71 ± 0.85	2015-01-01 ± 977.82d	
				Total	**39**	**20,139**			
			Con	Dev	4	1,968	Question Vote: 7.75 ± 3.96 Answer Vote: 10.50 ± 6.80	2012-11-13 ± 2.70d	
				Test	35	18,399	Question Vote: 7.40 ± 5.48 Answer Vote: 10.23 ± 7.54	2012-12-01 ± 6.52d	
				Total	**39**	**20,367**			
		Stack Overflow	Uncon	Dev	40	19,733	Question Vote: 3.45 ± 2.14 Answer Vote: 5.15 ± 2.14	2014-12-01 ± 1356.23d	0.06698
				Test	360	177,269	Question Vote: 3.45 ± 2.13 Answer Vote: 5.15 ± 2.11	2014-12-01 ± 1355.68d	
				Total	**400**	**197,002**			
			Con	Dev	40	19,932	Question Vote: 18.50 ± 41.99 Answer Vote: 26.68 ± 59.07	2013-06-25 ± 1119.50d	
				Test	360	177,118	Question Vote: 20.34 ± 42.03 Answer Vote: 28.37 ± 59.89	2013-06-11 ± 1119.19d	
				Total	**400**	**197,050**			

Table 28: Detailed per-subset statistics for Stack Exchange in the mid-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Domain	Model	Subset	Split	Partition	Samples	Tokens	Quality	Temporal	Lexical
Post-training	SFT	1B, 32B	Aya	Uncon	Dev	14	5,293	—	—	0.27350
					Test	140	47,321	—	—	
					Total	**154**	**52,614**			
				Con	Dev	14	4,910	—	—	
					Test	140	47,391	—	—	
					Total	**154**	**52,301**			
			Wildchat	Uncon	Dev	7	16,404	—	—	0.26453
					Test	64	148,225	—	—	
					Total	**71**	**164,629**			
				Con	Dev	7	16,388	—	—	
					Test	64	148,120	—	—	
					Total	**71**	**164,508**			
		7B, 13B	Aya	Uncon	Dev	13	5,117	—	—	0.43120
					Test	82	46,860	—	—	
					Total	**95**	**51,977**			
				Con	Dev	13	5,224	—	—	
					Test	82	47,024	—	—	
					Total	**95**	**52,248**			
			Wildchat	Uncon	Dev	7	16,523	—	—	0.28384
					Test	66	148,377	—	—	
					Total	**73**	**164,900**			
				Con	Dev	7	16,492	—	—	
					Test	66	148,578	—	—	
					Total	**73**	**165,070**			

Table 29: Detailed per-subset statistics for SFT in the post-training stage of OLMO-DETECT (SHIFTED). **Uncon** and **Con** denote the uncontaminated and contaminated splits, respectively. **Dev** refers to the development set used for hyperparameter tuning, while **Test** refers to the test set used for method evaluation.

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DCLM-Baseline	1B	0.5859	0.5997	**0.6009**	0.5332	0.5112	0.5588	0.4451	0.5673	0.5958	0.4720	0.5234	0.5364	0.4849	0.5388	0.5395
	7B	0.6333	**0.6450**	0.6258	0.5777	0.5624	0.6285	0.4813	0.6264	0.6409	0.4760	0.5402	0.3844	0.5149	0.5050	0.5601
	13B	0.6422	0.6582	0.6542	0.6119	0.5626	0.6275	0.5722	0.6282	**0.6677**	0.4772	0.5257	0.3494	0.5263	0.4867	0.5707
	32B	0.6544	0.6734	0.6728	0.6342	0.6396	0.6441	0.5545	0.6222	**0.6772**	0.4793	0.5030	0.4769	0.5342	0.4621	0.5877
	Model-Avg	0.6290	0.6441	0.6384	0.5893	0.5689	0.6147	0.5133	0.6110	**0.6454**	0.4761	0.5231	0.4368	0.5151	0.4981	0.5645

Table 30: Per-method AUC on OLMO-DETECT for the DCLM-Baseline domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
peS2o	1B	0.3498	0.6063	0.5134	0.4675	0.5097	0.5358	0.6441	0.6703	0.4766	0.5044	**0.7437**	0.5781	0.4800	0.4835	0.5402
	7B	0.3552	0.5916	0.5238	0.4198	0.5234	0.4785	0.5898	0.6873	0.4366	0.5406	**0.7008**	0.5487	0.4981	0.4935	0.5277
	13B	0.3568	0.5865	0.5712	0.4227	0.5118	0.4750	0.6454	0.6635	0.4746	0.5604	**0.7048**	0.5617	0.4969	0.3944	0.5304
	32B	0.3652	0.5892	0.5705	0.4039	0.4901	0.4747	0.5681	0.6643	0.4790	0.5700	**0.7131**	0.5490	0.5110	0.4078	0.5254
	Model-Avg	0.3568	0.5934	0.5447	0.4285	0.5088	0.4910	0.6119	0.6714	0.4667	0.5438	**0.7156**	0.5594	0.4965	0.4448	0.5309

Table 31: Per-method AUC on OLMO-DETECT for the peS2o domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	1B	0.7901	0.8111	0.8081	0.6703	0.5671	0.7479	0.6973	0.8209	**0.8520**	0.6944	0.7438	0.6412	0.6428	0.6460	0.7238
	7B	0.8015	0.8406	0.8026	0.7597	0.5806	**0.8788**	0.6579	0.8386	0.8732	0.6681	0.6605	0.6121	0.6476	0.6575	0.7342
	13B	0.8075	0.8494	0.7867	0.7759	0.5999	0.8691	0.6358	0.8452	**0.8719**	0.6743	0.7185	0.6349	0.7022	0.7258	0.7498
	32B	0.8215	0.8591	0.7997	0.8137	0.6255	**0.8824**	0.8250	0.8345	0.8670	0.6524	0.8311	0.4149	0.5294	0.7595	0.7511
	Model-Avg	0.8052	0.8400	0.7993	0.7549	0.5933	0.8446	0.7040	0.8348	**0.8660**	0.6723	0.7385	0.5758	0.6305	0.6972	0.7397

Table 32: Per-method AUC on OLMo-Detect for the StarCoder domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
OpenWebMath	1B	0.5059	0.5728	**0.6168**	0.5375	0.5223	0.5981	0.4958	0.5251	0.5343	0.5463	0.5440	0.5446	0.4753	0.4871	0.5361
	7B	0.5119	0.5835	**0.5865**	0.5364	0.5167	0.5692	0.4850	0.4730	0.5610	0.5459	0.5277	0.5343	0.4526	0.4514	0.5239
	13B	0.5212	0.5987	0.6121	0.5460	0.5139	**0.6209**	0.5039	0.4436	0.5752	0.5549	0.5193	0.5260	0.4922	0.4713	0.5357
	32B	0.5609	**0.6437**	0.5901	0.5825	0.6214	0.6364	0.4902	0.4328	0.5867	0.5586	0.5956	0.5457	0.4982	0.4548	0.5570
	Model-Avg	0.5250	0.5997	0.6014	0.5506	0.5436	**0.6061**	0.4937	0.4686	0.5643	0.5514	0.5466	0.5376	0.4796	0.4662	0.5382

Table 33: Per-method AUC on OLMo-Detect for the OpenWebMath domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
GSM8K	1B	0.7309	**0.7599**	0.7217	0.6919	0.5762	0.6913	0.3232	0.7508	0.7190	0.6344	0.5018	0.4696	0.5255	0.4803	0.6126
	7B	0.8483	0.8875	0.8257	0.8415	0.5288	**0.9055**	0.0556	0.8884	0.8282	0.7227	0.5144	0.5059	0.5500	0.5152	0.6727
	13B	0.9848	**0.9915**	0.9311	0.9734	0.5895	0.9852	0.0158	0.9899	0.9608	0.8849	0.5441	0.4933	0.6316	0.5324	0.7506
	32B	0.9798	0.9899	0.9296	0.9650	0.5057	0.9819	0.0142	**0.9922**	0.9211	0.9095	0.5143	0.5130	0.7018	0.5452	0.7474
	Model-Avg	0.8860	**0.9072**	0.8520	0.8679	0.5501	0.8910	0.1022	0.9053	0.8573	0.7879	0.5187	0.4955	0.6022	0.5183	0.6958

Table 34: Per-method AUC on OLMo-Detect for the GSM8K domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	1B	0.5732	**0.6373**	0.5437	0.5661	0.5551	0.5780	0.5138	0.6189	0.6217	0.5549	0.6195	0.4959	0.5001	0.4694	0.5605
	7B	0.6097	**0.6950**	0.5850	0.6454	0.6514	0.6512	0.6768	0.6036	0.6590	0.5561	0.5973	0.5144	0.5122	0.4456	0.6002
	13B	0.6203	0.7213	0.6035	0.6486	0.6758	0.6962	**0.7345**	0.6536	0.6740	0.5655	0.6547	0.5652	0.5267	0.5130	0.6323
	32B	0.6459	0.7352	0.5954	0.6729	0.6407	0.7068	**0.7739**	0.6639	0.7266	0.5767	0.7412	0.7238	0.5113	0.5095	0.6588
	Model-Avg	0.6123	**0.6972**	0.5819	0.6332	0.6308	0.6581	0.6747	0.6350	0.6703	0.5633	0.6532	0.5748	0.5126	0.4844	0.6130

Table 35: Per-method AUC on OLMo-Detect for the Stack Exchange domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	1B	0.5893	0.6121	0.4856	0.5674	**0.7026**	0.5723	0.5218	0.5406	0.5157	0.5199	0.5654	0.4377	0.5718	0.4479	0.5464
	7B	0.7521	0.7167	0.5694	0.6919	**0.7871**	0.7095	0.7111	0.7317	0.6342	0.6536	0.5899	0.5362	0.5241	0.4242	0.6451
	13B	0.6903	0.6197	0.5104	0.6716	**0.7177**	0.6272	0.6246	0.6371	0.6072	0.6179	0.5535	0.4978	0.5450	0.5033	0.6017
	32B	0.5865	0.5966	0.5090	0.5802	**0.6684**	0.5823	0.6155	0.5241	0.5197	0.5169	0.5487	0.5134	0.5425	0.4354	0.5528
	Model-Avg	0.6545	0.6363	0.5186	0.6278	**0.7189**	0.6228	0.6182	0.6084	0.5692	0.5771	0.5644	0.4963	0.5458	0.4527	0.5865

Table 36: Per-method AUC on OLMo-Detect for the SFT domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DPO	1B	0.4425	0.5284	**0.6443**	0.4553	0.4731	0.4535	0.4840	0.5769	0.6041	0.5280	0.5602	0.4592	0.5645	0.4973	0.5194
	7B	0.4099	0.4342	0.5483	0.3998	0.3857	0.4022	0.5934	**0.6276**	0.4864	0.4900	0.5653	0.4783	0.5245	0.4519	0.4855
	13B	0.3853	0.4708	0.6205	0.3858	0.4443	0.3938	**0.6637**	0.6475	0.5255	0.4816	0.3091	0.4639	0.5992	0.4271	0.4870
	32B	0.4155	0.4871	0.5893	0.4185	0.4597	0.4025	**0.7228**	0.6520	0.5890	0.5538	0.2529	0.4946	0.5861	0.4432	0.5048
	Model-Avg	0.4133	0.4801	0.6006	0.4148	0.4407	0.4130	0.6160	**0.6260**	0.5513	0.5133	0.4219	0.4740	0.5686	0.4549	0.4992

Table 37: Per-method AUC on OLMo-Detect for the DPO domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
RLVR	1B	0.5546	0.5788	0.5856	0.5493	0.4756	0.5149	0.5011	0.6164	**0.6277**	0.5138	0.5633	0.4836	0.5675	0.5071	0.5457
	7B	0.5732	0.5833	0.5849	0.5474	0.4640	0.5361	0.5907	**0.7463**	0.6372	0.5100	0.5149	0.4247	0.5284	0.5109	0.5537
	13B	0.6139	0.6323	0.6161	0.5783	0.4974	0.5736	0.4220	**0.6960**	0.6438	0.5434	0.5401	0.3763	0.5500	0.4817	0.5546
	32B	0.6010	0.6207	0.6028	0.5842	0.4637	0.6162	0.5496	**0.6712**	0.6428	0.5409	0.5447	0.4773	0.5481	0.4626	0.5661
	Model-Avg	0.5857	0.6038	0.5974	0.5648	0.4752	0.5602	0.5159	**0.6825**	0.6379	0.5270	0.5408	0.4405	0.5485	0.4906	0.5550

Table 38: Per-method AUC on OLMo-Detect for the RLVR domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	Assembly	1B	0.5833	0.5046	0.3843	0.4136	0.5602	0.3302	0.3611	0.5247	0.5216	0.6451	**0.8920**	0.4706	0.5077	0.5741	0.5195
		7B	0.5540	0.5340	0.4969	0.4506	0.6049	0.5062	0.4167	0.5617	0.5679	**0.6944**	0.5648	0.5458	0.5015	0.5185	0.5370
		13B	0.5617	0.5062	0.4630	0.4861	0.5170	0.4877	0.4136	0.5247	0.5741	**0.7068**	0.6944	0.4967	0.5926	0.6667	0.5494
		32B	0.5710	0.5340	0.4583	0.5355	0.5988	0.5185	0.4537	0.5448	0.6080	0.6914	**0.8580**	0.4314	0.6358	0.6127	0.5751
		Model-Avg	0.5675	0.5197	0.4506	0.4714	0.5702	0.4607	0.4113	0.5390	0.5679	0.6844	**0.7523**	0.4861	0.5594	0.5930	0.5453
	Java	1B	0.7994	0.8466	0.8320	0.6881	0.5678	0.7787	0.7108	0.8489	**0.8751**	0.7010	0.5956	0.6562	0.6501	0.6509	0.7287
		7B	0.8130	0.8781	0.8202	0.7722	0.5855	**0.9029**	0.6718	0.8547	0.8948	0.6672	0.7563	0.6166	0.6553	0.6654	0.7539
		13B	0.8213	0.8882	0.8045	0.7872	0.6076	**0.8945**	0.6476	0.8716	0.8927	0.6740	0.7425	0.6421	0.7082	0.7287	0.7651
		32B	0.8354	0.8872	0.8148	0.8272	0.6318	**0.9080**	0.8419	0.8635	0.8877	0.6516	0.8041	0.4122	0.5208	0.7665	0.7609
		Model-Avg	0.8173	0.8750	0.8179	0.7687	0.5982	0.8710	0.7180	0.8597	**0.8876**	0.6734	0.7246	0.5818	0.6336	0.7029	0.7521

Table 39: Per-method AUC on OLMo-Detect for the StarCoder domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	Math Stack Exchange	1B	0.6065	0.6531	0.4955	0.5755	0.5033	0.5755	0.6792	0.6184	0.6661	0.5273	**0.7388**	0.5153	0.6110	0.6498	0.6011
		7B	0.6249	0.6751	0.5759	0.6800	0.6278	0.6253	0.5200	0.6441	**0.7224**	0.5380	0.7016	0.4171	0.5090	0.4629	0.5946
		13B	0.6371	0.7135	0.5690	0.6947	0.6012	0.6963	0.6482	0.6482	**0.7510**	0.5437	0.6931	0.4960	0.4449	0.5127	0.6178
		32B	0.6829	0.7494	0.5437	0.7131	0.7008	0.6824	0.7755	0.6735	0.7837	0.5682	**0.8518**	0.7134	0.3927	0.5784	0.6721
		Model-Avg	0.6379	0.6978	0.5460	0.6658	0.6083	0.6449	0.6557	0.6461	0.7308	0.5443	**0.7463**	0.5354	0.4894	0.5510	0.6214
	Stack Overflow	1B	0.5768	**0.6356**	0.5480	0.5647	0.5605	0.5806	0.5045	0.6219	0.6192	0.5570	0.5003	0.4949	0.4895	0.4501	0.5503
		7B	0.6151	**0.6980**	0.5860	0.6462	0.6623	0.6595	0.6956	0.6064	0.6546	0.5582	0.4931	0.5214	0.5128	0.4425	0.5965
		13B	0.6277	0.7225	0.6067	0.6534	0.6844	0.7045	**0.7476**	0.6590	0.6692	0.5674	0.6163	0.5726	0.5335	0.5133	0.6341
		32B	0.6583	0.7356	0.5996	0.6862	0.6341	0.7177	**0.7905**	0.6752	0.7236	0.5783	0.6305	0.7255	0.5211	0.5027	0.6556
		Model-Avg	0.6195	**0.6979**	0.5851	0.6376	0.6353	0.6656	0.6845	0.6406	0.6666	0.5652	0.5600	0.5786	0.5142	0.4772	0.6091

Table 40: Per-method AUC on OLMo-Detect for the Stack Exchange domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	Aya	1B	0.5436	0.5506	0.4896	0.5018	**0.6373**	0.5031	0.4558	0.5861	0.4859	0.4934	0.5811	0.4183	0.6015	0.4148	0.5188
		7B	0.7345	0.6066	0.5135	0.6517	**0.7674**	0.6481	0.7063	0.7277	0.5803	0.6689	0.5872	0.5785	0.5338	0.4366	0.6244
		13B	**0.6773**	0.5107	0.4521	0.6447	0.6504	0.5473	0.6065	0.6166	0.5348	0.6002	0.5094	0.5028	0.5551	0.4461	0.5610
		32B	0.5512	0.5406	0.5175	0.5367	0.6315	0.5443	**0.6447**	0.4834	0.4882	0.4833	0.4351	0.5223	0.5423	0.3777	0.5213
		Model-Avg	0.6267	0.5521	0.4932	0.5837	**0.6716**	0.5607	0.6033	0.6035	0.5223	0.5615	0.5282	0.5055	0.5582	0.4188	0.5564
	WildChat	1B	0.7179	0.7725	0.5031	0.7385	**0.8694**	0.7415	0.6880	0.4622	0.6289	0.5828	0.5498	0.4496	0.5062	0.4936	0.6217
		7B	0.7894	**0.8799**	0.6899	0.7485	0.8511	0.7984	0.7181	0.7727	0.7316	0.6434	0.5927	0.4670	0.5129	0.4127	0.6863
		13B	0.7223	0.8147	0.6312	0.7076	**0.8272**	0.7328	0.6455	0.7101	0.7082	0.6323	0.5976	0.5006	0.5302	0.5849	0.6675
		32B	0.7261	0.7629	0.5139	0.7643	**0.7941**	0.7075	0.6445	0.6523	0.6028	0.5918	0.6624	0.4685	0.5431	0.5466	0.6415
		Model-Avg	0.7389	0.8075	0.5845	0.7397	**0.8355**	0.7450	0.6740	0.6493	0.6679	0.6126	0.6006	0.4714	0.5231	0.5094	0.6543

Table 41: Per-method AUC on OLMO-DETECT for the SFT domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
RLVR	CoT-GSM8K	1B	0.8474	0.8897	0.7242	0.8306	0.6884	0.7800	**0.9747**	0.6769	0.7589	0.5000	0.5000	0.5960	0.4941	0.5743	0.7025
		7B	0.9850	0.9837	0.9112	0.9688	0.8232	0.9691	**0.9997**	0.7646	0.8736	0.5000	0.5000	0.4343	0.5059	0.4243	0.7602
		13B	**1.0000**	0.9963	0.9079	0.9813	0.7338	0.9880	0.0223	0.5431	0.7485	0.5000	0.5000	0.7879	0.4964	0.4067	0.6866
		32B	**1.0000**	0.9988	0.9635	0.9999	0.8754	0.9933	1.0000	0.5284	0.8725	0.5000	0.5000	0.5354	0.4895	0.4092	0.7619
		Model-Avg	0.9581	**0.9671**	0.8767	0.9451	0.7802	0.9326	0.7492	0.6282	0.8134	0.5000	0.5000	0.5884	0.4965	0.4536	0.7278
	CoT-MATH	1B	0.7364	0.6687	0.5710	**0.7847**	0.6424	0.7776	0.4173	0.7025	0.6731	0.6535	0.5000	0.5043	0.5486	0.5594	0.6242
		7B	0.8219	0.7653	0.5119	**0.8314**	0.6550	0.7758	0.4683	0.7947	0.6484	0.6463	0.5000	0.6271	0.5197	0.5645	0.6522
		13B	**0.9232**	0.8854	0.6792	0.9210	0.6376	0.8763	0.4236	0.9032	0.7849	0.6988	0.5000	0.4921	0.5632	0.4273	0.6940
		32B	**0.9259**	0.8587	0.7778	0.9258	0.6945	0.9237	0.2932	0.8641	0.7829	0.6768	0.5000	0.5987	0.5094	0.3887	0.6943
		Model-Avg	0.8518	0.7945	0.6350	**0.8657**	0.6574	0.8383	0.4006	0.8161	0.7223	0.6688	0.5000	0.5555	0.5352	0.4850	0.6662
	IFEval	1B	0.4713	0.4936	**0.6878**	0.4391	0.3516	0.4109	0.4209	0.6484	0.6095	0.5282	0.6266	0.4448	0.6249	0.4901	0.5177
		7B	0.4933	0.5277	0.6769	0.4332	0.3467	0.4413	**0.7479**	0.6999	0.6002	0.5134	0.5298	0.3607	0.5599	0.5324	0.5331
		13B	0.5446	0.6389	0.7222	0.5091	0.4428	0.5254	0.4164	**0.7337**	0.6573	0.5699	0.5801	0.3082	0.5883	0.4797	0.5512
		32B	0.5222	0.5901	0.7082	0.4798	0.3875	0.5835	0.5142	**0.7230**	0.6510	0.5531	0.5894	0.4557	0.5907	0.4461	0.5567
		Model-Avg	0.5079	0.5626	0.6988	0.4653	0.3821	0.4903	0.5249	**0.7012**	0.6295	0.5412	0.5815	0.3923	0.5909	0.4871	0.5397

Table 42: Per-method AUC on OLMO-DETECT for the RLVR domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
All Stages	1B	0.1326	0.1490	0.1413	0.1050	0.0735	0.0831	0.0343	**0.1629**	0.1381	0.0950	0.0750	0.0730	0.0660	0.0509	0.0985
	7B	0.2256	**0.2756**	0.1722	0.2123	0.0752	0.2211	0.0841	0.2218	0.2214	0.0911	0.1344	0.0624	0.0620	0.0610	0.1514
	13B	0.3713	0.4021	0.2266	0.2894	0.1065	0.2946	0.0277	**0.4447**	0.3269	0.2007	0.1275	0.0575	0.1043	0.0764	0.2183
	32B	0.3407	0.3651	0.2813	0.3345	0.0932	0.3791	0.1310	**0.4043**	0.2907	0.1713	0.1472	0.0750	0.1354	0.0832	0.2309
	Model-Avg	0.2676	0.2979	0.2054	0.2353	0.0871	0.2445	0.0693	**0.3084**	0.2443	0.1395	0.1210	0.0670	0.0919	0.0679	0.1748

Table 43: Per-method TPR@5%FPR on OLMO-DETECT for all training stages. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Pre-training	1B	0.0549	0.1718	**0.2155**	0.0624	0.0416	0.0704	0.0489	0.1551	0.2017	0.1026	0.0719	0.0931	0.0630	0.1432	0.1069
	7B	0.0800	0.2232	0.2080	0.0903	0.0517	0.1647	0.0274	0.1325	**0.2434**	0.0609	0.1988	0.0874	0.0560	0.1480	0.1266
	13B	0.1131	**0.2411**	0.1818	0.0885	0.0716	0.1611	0.0334	0.1945	0.2351	0.0716	0.1981	0.0616	0.0686	0.1110	0.1308
	32B	0.1062	0.2220	0.1794	0.1048	0.0998	0.1862	0.1229	0.1205	**0.2351**	0.0644	0.1968	0.0673	0.0443	0.1347	0.1346
	Model-Avg	0.0886	0.2145	0.1962	0.0865	0.0662	0.1456	0.0582	0.1507	**0.2288**	0.0749	0.1664	0.0774	0.0580	0.1342	0.1247

Table 44: Per-method TPR@5%FPR on OLMO-DETECT for the pre-training stage. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Mid-training	1B	0.1761	**0.2057**	0.1014	0.1568	0.0742	0.1074	0.0255	0.1794	0.1712	0.0878	0.0900	0.0373	0.0685	0.0353	0.1083
	7B	0.3496	0.4294	0.1691	0.3588	0.1281	**0.4752**	0.0090	0.3589	0.3116	0.1366	0.1020	0.0544	0.0683	0.0450	0.2140
	13B	0.6661	0.6944	0.3943	0.6314	0.1401	0.7042	0.0000	**0.7155**	0.6006	0.3934	0.0689	0.0533	0.1195	0.0755	0.3755
	32B	0.6536	0.6847	0.4664	0.6168	0.1462	0.6899	0.1051	**0.7237**	0.5120	0.3904	0.1975	0.1183	0.1829	0.0883	0.3983
	Model-Avg	0.4613	**0.5035**	0.2828	0.4410	0.1221	0.4942	0.0349	0.4944	0.3988	0.2520	0.1146	0.0658	0.1098	0.0610	0.2740

Table 45: Per-method TPR@5%FPR on OLMO-DETECT for the mid-training stage. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Post-training	1B	0.0914	0.0708	**0.1306**	0.1018	0.1017	0.0811	0.0901	0.1287	0.1030	0.0759	0.0690	0.0291	0.0557	0.0383	0.0834
	7B	0.1344	0.1160	0.1490	0.1219	0.1389	0.1075	**0.2178**	0.1542	0.0905	0.0948	0.0701	0.0323	0.0514	0.0501	0.1092
	13B	0.1070	0.0839	0.1757	0.1016	0.1220	0.0825	0.1469	**0.1888**	0.1245	0.1273	0.0726	0.0391	0.0556	0.0405	0.1049
	32B	0.1396	0.0920	**0.1565**	0.0846	0.0598	0.0539	0.1524	0.1327	0.0880	0.0959	0.0474	0.0703	0.0817	0.0352	0.0921
	Model-Avg	0.1181	0.0907	**0.1530**	0.1025	0.1056	0.0813	0.1518	0.1511	0.1015	0.0985	0.0648	0.0427	0.0611	0.0410	0.0974

Table 46: Per-method TPR@5%FPR on OLMO-DETECT for the post-training stage. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DCLM-Baseline	1B	0.0896	0.0672	**0.1506**	0.1470	0.0201	0.1269	0.0299	0.0672	0.1418	0.0672	0.0597	0.0533	0.0266	0.0672	0.0796
	7B	0.1545	0.1119	0.1716	0.1659	0.0575	0.1866	0.0448	**0.2090**	0.1866	0.0522	0.0672	0.0400	0.0786	0.0448	0.1122
	13B	0.1716	0.1493	0.1496	0.1716	0.0873	0.1418	0.0522	0.2388	**0.2612**	0.0746	0.0821	0.0133	0.1004	0.0149	0.1221
	32B	0.2164	0.2090	0.1343	0.2164	0.1343	0.1940	0.0597	0.2761	**0.3358**	0.0821	0.1194	0.0267	0.1158	0.0448	0.1546
	Model-Avg	0.1580	0.1343	0.1515	0.1752	0.0748	0.1623	0.0467	0.1978	**0.2314**	0.0690	0.0821	0.0333	0.0804	0.0429	0.1171

Table 47: Per-method TPR@5%FPR on OLMO-DETECT for the DCLM-Baseline domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
peS2o	1B	0.0373	0.1037	0.0623	0.0502	0.1120	0.0705	0.1701	**0.2407**	0.0539	0.0456	0.0250	0.0750	0.0425	0.0415	0.0807
	7B	0.0456	0.1203	0.0626	0.0456	0.0788	0.0373	0.1162	0.2116	0.0498	0.0581	**0.4950**	0.1200	0.0459	0.0415	0.1092
	13B	0.0539	0.1037	0.1203	0.0456	0.1039	0.0539	0.1992	0.2905	0.0456	0.0788	**0.4400**	0.0850	0.0570	0.0415	0.1228
	32B	0.0539	0.1369	0.1362	0.0603	0.0830	0.0539	0.1120	0.2863	0.0705	0.0913	**0.4500**	0.1450	0.0893	0.0830	0.1323
	Model-Avg	0.0477	0.1162	0.0954	0.0504	0.0944	0.0539	0.1494	0.2573	0.0549	0.0684	**0.3525**	0.1062	0.0587	0.0519	0.1112

Table 48: Per-method TPR@5%FPR on OLMO-DETECT for the peS2o domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	1B	0.0459	**0.4281**	0.2763	0.0513	0.0489	0.0520	0.0642	0.0856	0.3761	0.0887	0.1161	0.1215	0.0786	0.2661	0.1500
	7B	0.0550	**0.5107**	0.3008	0.0836	0.0489	0.5076	0.0917	0.1284	0.5076	0.0642	0.1450	0.1402	0.0439	0.1927	0.2014
	13B	0.0673	**0.5209**	0.2804	0.0734	0.0520	0.4709	0.0612	0.1590	0.4862	0.0917	0.1231	0.0218	0.0673	0.1896	0.1903
	32B	0.0703	**0.5963**	0.2263	0.0755	0.0856	0.5168	0.1223	0.1040	0.4557	0.0520	0.1150	0.0062	0.0411	0.2630	0.1950
	Model-Avg	0.0596	**0.5140**	0.2710	0.0709	0.0588	0.3868	0.0848	0.1192	0.4564	0.0741	0.1248	0.0724	0.0577	0.2278	0.1842

Table 49: Per-method TPR@5%FPR on OLMO-DETECT for the StarCoder domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
OpenWebMath	1B	0.0441	0.0735	0.1029	0.0713	**0.1196**	0.1029	0.0147	0.0441	0.0809	0.0809	0.0868	0.0980	0.0441	0.0441	0.0720
	7B	0.0441	0.0735	0.0794	**0.1029**	0.0882	0.0956	0.0221	0.0294	0.0074	0.0809	0.0882	0.0784	0.0221	0.0294	0.0601
	13B	0.0515	0.0956	0.0515	0.0588	0.0956	0.0662	0.0221	0.0294	0.0294	0.0662	**0.1471**	0.0980	0.0294	0.0147	0.0611
	32B	0.0779	**0.1618**	0.0441	0.0588	0.1382	0.1324	0.0588	0.0294	0.0882	0.0882	0.1029	0.0490	0.0721	0.0221	0.0803
	Model-Avg	0.0544	0.1011	0.0695	0.0730	**0.1104**	0.0993	0.0294	0.0331	0.0515	0.0790	0.1063	0.0809	0.0419	0.0276	0.0684

Table 50: Per-method TPR@5%FPR on OLMO-DETECT for the OpenWebMath domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (<0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
GSM8K	1B	0.2150	**0.2529**	0.1735	0.1998	0.1427	0.1227	0.0309	0.2348	0.1622	0.0832	0.0550	0.0398	0.0717	0.0277	0.1294
	7B	0.4296	0.5368	0.2141	0.4677	0.1354	**0.5977**	0.0000	0.4888	0.3682	0.2028	0.0491	0.0382	0.0745	0.0470	0.2607
	13B	0.9196	**0.9680**	0.5363	0.8809	0.2174	0.9424	0.0000	0.9594	0.7801	0.5422	0.0250	0.0443	0.1184	0.0726	0.5005
	32B	0.8937	0.9605	0.6105	0.8429	0.1438	0.9178	0.0011	**0.9616**	0.6681	0.6009	0.0650	0.0291	0.1930	0.0980	0.4990
	Model-Avg	0.6145	**0.6795**	0.3836	0.5978	0.1598	0.6452	0.0080	0.6612	0.4947	0.3573	0.0485	0.0379	0.1144	0.0613	0.3474

Table 51: Per-method TPR@5%FPR on OLMO-DETECT for the GSM8K domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (<0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	1B	0.0633	**0.1544**	0.0809	0.0462	0.0858	0.0608	0.0354	0.0911	0.0937	0.0886	0.1250	0.0387	0.0478	0.0380	0.0750
	7B	0.0652	**0.2253**	0.0991	0.0848	0.1657	0.1468	0.0810	0.0962	0.1342	0.0861	0.1550	0.0599	0.0525	0.0304	0.1059
	13B	0.0658	**0.2481**	0.1135	0.1057	0.1753	0.1848	0.1316	0.1165	0.1873	0.0911	0.1129	0.0775	0.0449	0.0329	0.1206
	32B	0.0706	0.2430	0.0986	0.0706	0.1228	0.1646	0.2481	0.1139	0.1696	0.0987	**0.3300**	0.2218	0.0506	0.0658	0.1478
	Model-Avg	0.0662	**0.2177**	0.0980	0.0768	0.1374	0.1393	0.1240	0.1044	0.1462	0.0911	0.1807	0.0995	0.0490	0.0418	0.1123

Table 52: Per-method TPR@5%FPR on OLMO-DETECT for the Stack Exchange domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (<0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	1B	0.1912	0.0931	0.1680	0.2216	**0.2951**	0.2010	0.1275	0.1029	0.1029	0.0784	0.0442	0.0191	0.1029	0.0372	0.1275
	7B	0.4392	0.3176	0.1568	0.4203	**0.4730**	0.3514	0.1284	0.3378	0.2162	0.1351	0.0654	0.0522	0.0730	0.0878	0.2324
	13B	0.2865	0.1757	0.0351	0.3378	**0.4257**	0.2432	0.0608	0.2568	0.1216	0.0946	0.0562	0.0348	0.0743	0.0743	0.1627
	32B	0.1912	0.0931	0.1176	**0.2412**	0.2255	0.1324	0.2206	0.1225	0.0882	0.0980	0.0748	0.0191	0.1225	0.0532	0.1286
	Model-Avg	0.2770	0.1699	0.1194	0.3052	**0.3548**	0.2320	0.1343	0.2050	0.1322	0.1015	0.0602	0.0313	0.0932	0.0631	0.1628

Table 53: Per-method TPR@5%FPR on OLMO-DETECT for the SFT domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (<0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DPO	1B	0.0745	0.0588	0.1480	0.0833	0.0441	0.0637	0.1471	0.1225	0.0882	0.0441	0.0980	0.0710	**0.1520**	0.0400	0.0882
	7B	0.0526	0.0526	0.1289	0.0632	0.0421	0.0579	**0.2105**	0.0158	0.0368	0.0105	0.0842	0.0692	0.0776	0.0276	0.0664
	13B	0.0253	0.0505	0.2626	0.0247	0.0303	0.0253	**0.3434**	0.1162	0.0556	0.0253	0.0808	0.0276	0.1813	0.0415	0.0922
	32B	0.0266	0.0266	0.1689	0.0160	0.0319	0.0053	**0.3032**	0.0957	0.0372	0.0532	0.0106	0.0667	0.1596	0.0330	0.0739
	Model-Avg	0.0447	0.0471	0.1771	0.0468	0.0371	0.0381	**0.2510**	0.0876	0.0544	0.0333	0.0684	0.0586	0.1426	0.0355	0.0802

Table 54: Per-method TPR@5%FPR on OLMO-DETECT for the DPO domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (<0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
		PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
RLVR	1B	0.2025	0.0407	0.1587	0.0446	0.0434	0.0271	0.0650	**0.2087**	0.1355	0.0894	0.0649	0.0579	0.0441	0.0650	0.0891
	7B	0.2074	0.0732	0.1801	0.1491	0.0515	0.0407	**0.2629**	0.1843	0.1816	0.0867	0.0606	0.0316	0.0608	0.0650	0.1168
	13B	0.2980	0.1274	0.2072	0.1416	0.0949	0.1003	0.0461	**0.3388**	0.2222	0.1680	0.0809	0.0526	0.0631	0.0623	0.1431
	32B	0.2779	0.1301	0.2017	0.2704	0.0512	0.2602	0.0596	**0.2873**	0.1870	0.1409	0.0569	0.1105	0.0729	0.0608	0.1548
	Model-Avg	0.2465	0.0929	0.1869	0.1514	0.0602	0.1071	0.1084	**0.2548**	0.1816	0.1212	0.0658	0.0631	0.0602	0.0633	0.1260

Table 55: Per-method TPR@5%FPR on OLMO-DETECT for the RLVR domain. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	Assembly	1B	**0.2778**	0.0556	0.0000	0.1111	0.0000	0.0000	0.0000	0.1111	0.1111	0.1667	0.2222	0.0000	0.0556	0.0000	0.0794
		7B	**0.2222**	0.1111	0.0000	0.1111	0.0000	0.1667	0.1111	0.0556	0.1667	0.2222	0.0000	0.1111	0.0556	0.0556	0.0992
		13B	0.0556	0.0000	0.0000	0.1667	0.1111	0.0556	0.1111	0.0000	**0.2222**	0.1667	0.1111	0.0556	0.2222	0.0000	0.0913
		32B	**0.2778**	0.0556	0.0556	0.1111	0.2778	0.0556	0.2222	0.0000	0.1667	0.1111	0.0500	0.0000	0.1111	0.1111	0.1147
		Model-Avg	**0.2084**	0.0556	0.0139	0.1250	0.0972	0.0695	0.1111	0.0417	0.1667	0.1667	0.0958	0.0417	0.1111	0.0417	0.0961
	Java	1B	0.0453	**0.4434**	0.2904	0.0501	0.0518	0.1133	0.0647	0.1650	0.3722	0.0939	0.0100	0.1254	0.0773	0.2460	0.1535
		7B	0.0583	**0.5405**	0.3154	0.0615	0.0500	0.4984	0.0615	0.1812	0.5178	0.0647	0.2900	0.1452	0.0403	0.1974	0.2159
		13B	0.0712	**0.5502**	0.2830	0.0759	0.0453	0.4919	0.0583	0.2945	0.5049	0.0971	0.1350	0.0198	0.0647	0.2006	0.2066
		32B	0.0744	**0.6181**	0.2345	0.0744	0.0841	0.5502	0.1197	0.1392	0.4725	0.0485	0.1800	0.0066	0.0361	0.2686	0.2076
		Model-Avg	0.0623	**0.5381**	0.2808	0.0655	0.0578	0.4134	0.0760	0.1950	0.4668	0.0761	0.1537	0.0742	0.0546	0.2281	0.1959

Table 56: Per-method TPR@5%FPR on OLMO-DETECT for the StarCoder domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	Math Stack Exchange	1B	0.0286	0.0571	0.1357	0.0857	0.0286	0.0286	0.0571	0.0571	0.1143	0.0571	**0.2000**	0.1111	0.2000	0.1143	0.0911
		7B	0.0000	0.0857	0.1536	0.0571	0.0857	0.1143	0.0286	0.0286	0.1429	0.0286	**0.2571**	0.0741	0.1357	0.0000	0.0851
		13B	0.0571	0.0571	0.1429	0.1071	0.1429	0.2000	0.0571	0.0857	**0.2857**	0.0571	0.0857	0.2593	0.0286	0.0286	0.1139
		32B	0.1143	0.0286	0.0857	0.1143	0.1143	0.2000	0.3143	0.1143	0.2571	0.0857	**0.4571**	0.3333	0.0286	0.0286	0.1626
		Model-Avg	0.0500	0.0571	0.1295	0.0910	0.0929	0.1357	0.1143	0.0714	0.2000	0.0571	**0.2500**	0.1944	0.0982	0.0429	0.1132
	Stack Overflow	1B	0.0694	**0.1750**	0.0875	0.0722	0.0889	0.0500	0.0389	0.1028	0.0972	0.0861	0.0500	0.0311	0.0444	0.0389	0.0737
		7B	0.0750	**0.2417**	0.1036	0.1028	0.1750	0.1500	0.0833	0.0972	0.1333	0.0944	0.0528	0.0623	0.0472	0.0333	0.1037
		13B	0.0778	**0.2444**	0.1176	0.1056	0.1778	0.1833	0.1333	0.1389	0.1833	0.1000	0.1400	0.0623	0.0472	0.0361	0.1248
		32B	0.0667	**0.2639**	0.1022	0.1046	0.1333	0.2111	0.2611	0.1306	0.1639	0.1083	0.2028	0.1790	0.0528	0.0528	0.1452
		Model-Avg	0.0722	**0.2313**	0.1027	0.0963	0.1437	0.1486	0.1291	0.1174	0.1444	0.0972	0.1114	0.0837	0.0479	0.0403	0.1119

Table 57: Per-method TPR@5%FPR on OLMO-DETECT for the Stack Exchange domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	Aya	1B	0.1071	0.0143	0.1429	0.1393	**0.1571**	0.0857	0.1071	0.1357	0.0429	0.0571	0.0571	0.0250	0.1357	0.0394	0.0890
		7B	**0.5000**	0.1829	0.1110	0.4024	0.3780	0.2317	0.2317	0.4268	0.1585	0.1098	0.0854	0.0278	0.1000	0.0854	0.2165
		13B	**0.3171**	0.0854	0.0366	0.2817	0.2695	0.1341	0.0610	0.2317	0.0610	0.0976	0.0366	0.0278	0.1829	0.0610	0.1346
		32B	0.1393	0.0286	0.1214	0.1214	0.1286	0.0643	**0.2571**	0.0714	0.0500	0.0786	0.0714	0.0000	0.1571	0.0630	0.0966
		Model-Avg	**0.2659**	0.0778	0.1030	0.2362	0.2333	0.1290	0.1642	0.2164	0.0781	0.0858	0.0626	0.0202	0.1439	0.0622	0.1342
	WildChat	1B	0.3125	0.2344	0.1444	0.2375	**0.6719**	0.2812	0.1406	0.0781	0.2812	0.1562	0.0312	0.0541	0.0469	0.0492	0.1942
		7B	0.4394	0.5000	0.2187	0.4611	**0.6515**	0.5455	0.0606	0.5000	0.3030	0.1667	0.0455	0.0698	0.0303	0.0909	0.2916
		13B	0.2576	0.3182	0.1061	0.2515	**0.6212**	0.3333	0.0606	0.3485	0.1364	0.1364	0.0758	0.0930	0.0455	0.0758	0.2043
		32B	0.2250	0.2188	0.1578	0.2219	**0.4375**	0.2188	0.0938	0.1719	0.1406	0.1719	0.0781	0.0541	0.0531	0.0656	0.1649
		Model-Avg	0.3086	0.3178	0.1568	0.2930	**0.5955**	0.3447	0.0889	0.2746	0.2153	0.1578	0.0576	0.0678	0.0440	0.0704	0.2138

Table 58: Per-method TPR@5%FPR on OLMO-DETECT for the SFT domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Domain	Subset	Model	Likelihood-based											Pert.	Out.-Dist.	Prompt	Method-Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
RLVR	GSM8K	1B	0.5080	0.5795	0.1909	0.4977	0.1827	0.1932	**0.9432**	0.2045	0.2955	0.0500	0.0500	0.0000	0.0379	0.1477	0.2772
		7B	0.9545	0.9091	0.5259	0.9110	0.3159	0.8864	**1.0000**	0.1136	0.4886	0.0500	0.0500	0.0000	0.0659	0.0455	0.4512
		13B	**1.0000**	0.9545	0.5164	0.8398	0.1136	0.9318	0.0000	0.2159	0.3182	0.0500	0.0500	0.4545	0.0471	0.0045	0.3926
		32B	**1.0000**	1.0000	0.8335	1.0000	0.3614	0.9659	1.0000	0.1040	0.5909	0.0500	0.0500	0.0909	0.0394	0.0045	0.5065
		Model-Avg	**0.8656**	0.8608	0.5167	0.8121	0.2434	0.7443	0.7358	0.1595	0.4233	0.0500	0.0500	0.1363	0.0476	0.0505	0.4069
	MATH	1B	0.2632	0.1828	0.0968	0.1849	0.0000	0.1575	0.0108	**0.2903**	0.2581	0.1290	0.0500	0.0769	0.0753	0.0430	0.1299
		7B	**0.4455**	0.2903	0.1185	0.3907	0.0680	0.3978	0.0108	0.3118	0.1828	0.1398	0.0500	0.0769	0.1145	0.0215	0.1871
		13B	**0.7704**	0.5699	0.1703	0.7113	0.0323	0.4731	0.0000	0.4409	0.2043	0.1935	0.0500	0.0513	0.0962	0.0108	0.2696
		32B	0.5860	0.4839	0.2949	0.6151	0.1608	**0.6452**	0.0000	0.2473	0.1828	0.1290	0.0500	0.0000	0.0500	0.0430	0.2491
		Model-Avg	**0.5163**	0.3817	0.1701	0.4755	0.0653	0.4184	0.0054	0.3226	0.2070	0.1478	0.0500	0.0513	0.0840	0.0296	0.2089
	IFEval	1B	0.1011	0.0426	0.2128	0.0319	0.0340	0.0213	0.0904	**0.3564**	0.1915	0.0745	0.0798	0.0500	0.1649	0.0266	0.1056
		7B	0.1117	0.0745	0.2739	0.0904	0.0638	0.0479	**0.3936**	0.2660	0.2128	0.0691	0.0713	0.0286	0.0851	0.0745	0.1331
		13B	0.2074	0.1436	0.3234	0.1298	0.1298	0.0851	0.0691	**0.4309**	0.3191	0.2021	0.1117	0.0500	0.1596	0.0957	0.1755
		32B	0.2011	0.2074	**0.3351**	0.1367	0.0638	0.1117	0.0638	0.3032	0.2926	0.1915	0.0638	0.1214	0.1106	0.0691	0.1623
		Model-Avg	0.1553	0.1170	0.2863	0.0972	0.0728	0.0665	0.1542	**0.3391**	0.2540	0.1343	0.0817	0.0625	0.1300	0.0665	0.1441

Table 59: Per-method TPR@5%FPR on OLMO-DETECT for the RLVR domain by subset. **Method-Avg** averages across all methods. The best-performing method for each row is highlighted in bold. Cell colors indicate detection strength: weak (< 0.10), moderate (0.10–0.30), and strong (≥ 0.30).

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Pre-training	1B	Matched	0.6321	0.6795	0.6650	0.5594	0.5423	0.6034	0.5784	0.7017	0.6738	0.5964	0.6387	0.5921	0.5430	0.5616	0.6120
		Shifted	0.7613	0.6579	0.5918	0.9985	0.2569	1.0000	0.6261	0.7318	0.7633	0.5312	0.7265	0.4944	0.6081	0.5170	0.6618
	7B	Matched	0.6477	0.7126	0.6595	0.6084	0.5552	0.7080	0.5690	0.7113	0.6910	0.5972	0.6073	0.5452	0.5523	0.5665	0.6237
		Shifted	0.7556	0.6487	0.5795	0.9849	0.5361	0.9971	0.5804	0.7282	0.8186	0.4677	0.4960	0.5509	0.7004	0.5334	0.6698
	13B	Matched	0.6508	0.7211	0.6698	0.6242	0.5620	0.6955	0.5923	0.6993	0.7042	0.6027	0.6171	0.5521	0.5806	0.5856	0.6327
		Shifted	0.7614	0.6703	0.5666	0.9542	0.6060	1.0000	0.6034	0.7282	0.7819	0.5383	0.7931	0.5904	0.6775	0.6085	0.7057
	32B	Matched	0.6675	0.7302	0.6705	0.6559	0.5739	0.7153	0.6653	0.6902	0.7071	0.6013	0.6607	0.4762	0.5272	0.5985	0.6386
		Shifted	0.7722	0.6773	0.5876	0.9576	0.5740	0.9979	0.7013	0.7393	0.7599	0.5378	0.6022	0.4452	0.5858	0.6312	0.6835
Avg Δ			+0.1131	−0.0473	−0.0848	+0.3618	−0.0651	+0.3182	+0.0266	+0.0312	+0.0869	−0.0806	+0.0235	−0.0212	+0.0922	−0.0055	+0.0535
***p*-value**			<0.01	0.01	<0.01	<0.01	0.45	<0.01	0.06	0.02	0.01	0.02	0.75	0.52	0.02	0.80	<0.01
Class.			I	D	D	I	R	I	I	I	I	D	R	R	I	R	I

Table 60: Comparison of AUC between OLMO-DETECT (**Matched**) and OLMO-DETECT (SHIFTED) (**Shifted**) on the pre-training stage. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Mid-training	1B	Matched	0.5732	0.6373	0.5437	0.5661	0.5551	0.5780	0.5138	0.6189	0.6217	0.5549	0.6195	0.4959	0.5001	0.4694	0.5605
		Shifted	0.5723	0.6796	0.5560	1.0000	0.2037	1.0000	0.5535	0.6186	0.7096	0.8166	0.6002	0.4743	0.4946	0.5149	0.6281
	7B	Matched	0.6097	0.6950	0.5850	0.6454	0.6514	0.6512	0.6768	0.6036	0.6590	0.5561	0.5973	0.5144	0.5122	0.4456	0.6002
		Shifted	0.6089	0.7283	0.5954	0.9956	0.6581	1.0000	0.6973	0.6620	0.7963	0.5584	0.6714	0.5148	0.5546	0.4785	0.6800
	13B	Matched	0.6203	0.7213	0.6035	0.6486	0.6758	0.6962	0.7345	0.6536	0.6740	0.5655	0.6547	0.5652	0.5267	0.5130	0.6323
		Shifted	0.6186	0.7519	0.6015	0.9642	0.6769	1.0000	0.7214	0.6955	0.7234	0.8276	0.5729	0.5746	0.5125	0.4981	0.6956
	32B	Matched	0.6459	0.7352	0.5954	0.6729	0.6407	0.7068	0.7739	0.6639	0.7266	0.5767	0.7412	0.7238	0.5113	0.5095	0.6588
		Shifted	0.6511	0.7832	0.5925	0.9752	0.6689	1.0000	0.8572	0.6463	0.7724	0.8289	0.7124	0.7297	0.4592	0.5159	0.7281
Avg Δ			+0.0005	+0.0386	+0.0045	+0.3505	−0.0789	+0.3419	+0.0326	+0.0206	+0.0801	+0.1946	−0.0140	−0.0015	−0.0073	+0.0175	+0.0700
***p*-value**			0.80	<0.01	0.35	<0.01	0.45	<0.01	0.20	0.33	0.03	0.06	0.70	0.85	0.73	0.29	<0.01
Class.			R	I	R	I	R	I	R	R	I	I	R	R	R	R	I

Table 61: Comparison of AUC between OLMO-DETECT (**Matched**) and OLMO-DETECT (SHIFTED) (**Shifted**) on the mid-training stage. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Post-training	1B	Matched	0.5183	0.5639	0.5518	0.5149	0.5901	0.5132	0.5110	0.5569	0.5637	0.5211	0.5628	0.4614	0.5689	0.4749	0.5338
		Shifted	0.5628	0.5657	0.5166	0.9916	0.3523	1.0000	0.5575	0.6154	0.6172	0.6671	0.6197	0.4736	0.5254	0.4817	0.6105
	7B	Matched	0.5576	0.5646	0.5570	0.5287	0.5628	0.5303	0.6519	0.6763	0.5430	0.5461	0.5776	0.5026	0.5243	0.4405	0.5545
		Shifted	0.5295	0.5647	0.5954	0.8891	0.5563	0.9887	0.6585	0.6264	0.6786	0.5228	0.6018	0.4744	0.6268	0.4539	0.6262
	13B	Matched	0.5090	0.5382	0.5743	0.5026	0.5626	0.4885	0.6493	0.6412	0.5546	0.5321	0.4313	0.4784	0.5759	0.4594	0.5355
		Shifted	0.4949	0.5192	0.5909	0.7658	0.5390	1.0000	0.5817	0.5215	0.6429	0.6205	0.4176	0.4829	0.5638	0.5155	0.5897
	32B	Matched	0.5000	0.5440	0.5439	0.5012	0.5647	0.4912	0.6532	0.5835	0.5579	0.5308	0.4008	0.5037	0.5633	0.4443	0.5273
		Shifted	0.5469	0.5443	0.4863	0.8547	0.5616	0.9949	0.6721	0.6228	0.6120	0.6365	0.6119	0.4918	0.4897	0.4624	0.6134
Avg Δ			+0.0123	−0.0042	−0.0095	+0.3635	−0.0677	+0.4901	+0.0011	−0.0180	+0.0829	+0.0792	+0.0696	−0.0059	−0.0067	+0.0236	+0.0722
***p*-value**			0.57	0.46	0.70	<0.01	0.32	<0.01	0.97	0.69	0.02	0.12	0.25	0.56	0.87	0.12	<0.01
Class.			R	R	R	I	R	I	R	R	I	R	R	R	R	R	I

Table 62: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the post-training stage. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DCLM-Baseline	1B	Matched	0.5859	0.5997	0.6009	0.5332	0.5112	0.5588	0.4451	0.5673	0.5958	0.4720	0.5234	0.5364	0.4849	0.5388	0.5395
		Shifted	0.5047	0.5600	0.5416	1.0000	0.0939	1.0000	0.3884	0.4477	0.5693	0.5718	0.5867	0.4684	0.4923	0.5818	0.5576
	7B	Matched	0.6333	0.6450	0.6258	0.5777	0.5624	0.6285	0.4813	0.6264	0.6409	0.4760	0.5402	0.3844	0.5149	0.5050	0.5601
		Shifted	0.5278	0.5847	0.5831	0.9827	0.5407	0.9931	0.4692	0.0000	0.7620	0.5695	0.5718	0.4064	0.5261	0.5611	0.5770
	13B	Matched	0.6422	0.6582	0.6542	0.6119	0.5626	0.6275	0.5722	0.6282	0.6677	0.4772	0.5257	0.3494	0.5263	0.4867	0.5707
		Shifted	0.5667	0.6187	0.6390	0.9160	0.5983	1.0000	0.5796	0.5723	0.6889	0.6391	0.5972	0.3872	0.5228	0.5156	0.6315
	32B	Matched	0.6544	0.6734	0.6728	0.6342	0.6396	0.6441	0.5545	0.6222	0.6772	0.4793	0.5030	0.4769	0.5342	0.4621	0.5877
		Shifted	0.6187	0.6508	0.6948	0.9506	0.6378	0.9928	0.5693	1.0000	0.7525	0.6393	0.5488	0.4888	0.5270	0.4911	0.6830
Avg Δ			−0.0745	−0.0405	−0.0238	+0.3731	−0.1013	+0.3818	−0.0116	−0.1060	+0.0478	+0.1288	+0.0530	+0.0009	+0.0020	+0.0392	+0.0478
***p*-value**			0.01	0.01	0.27	<0.01	0.41	<0.01	0.52	0.64	0.23	<0.01	<0.01	0.97	0.68	<0.01	0.09
Class.			D	D	R	I	R	I	R	R	R	I	I	R	R	I	I

Table 63: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the DCLM-Baseline domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
peS2o	1B	Matched	0.3498	0.6063	0.5134	0.4675	0.5097	0.5358	0.6441	0.6703	0.4766	0.5044	0.2554	0.5781	0.4800	0.4835	0.5053
		Shifted	0.6021	0.3208	0.5893	1.0000	0.1161	1.0000	0.7942	0.5226	0.6741	0.3203	0.9434	0.5517	0.5066	0.4473	0.5992
	7B	Matched	0.3552	0.5916	0.5238	0.4198	0.5234	0.4785	0.5898	0.6873	0.4366	0.5406	0.7046	0.5487	0.4981	0.4935	0.5280
		Shifted	0.5902	0.3549	0.6455	0.9984	0.4958	1.0000	0.6768	0.0000	0.8184	0.2674	0.9008	0.7399	0.6151	0.4242	0.6091
	13B	Matched	0.3568	0.5865	0.5712	0.4227	0.5118	0.4750	0.6454	0.6635	0.4746	0.5604	0.6997	0.5617	0.4969	0.3944	0.5300
		Shifted	0.6060	0.3904	0.5790	0.9617	0.5284	1.0000	0.8404	0.6476	0.7158	0.3061	0.8559	0.7210	0.5857	0.4901	0.6592
	32B	Matched	0.3652	0.5892	0.5705	0.4039	0.4901	0.4747	0.5681	0.6643	0.4790	0.5700	0.7131	0.5490	0.5110	0.4078	0.5254
		Shifted	0.6029	0.3937	0.5572	0.9763	0.5450	1.0000	0.7188	1.0000	0.6906	0.3062	0.1299	0.4963	0.5265	0.4158	0.5971
Avg Δ			+0.2435	−0.2284	+0.0480	+0.5556	−0.0874	+0.5090	+0.1457	−0.1288	+0.2580	−0.2438	+0.1143	+0.0679	+0.0620	−0.0004	+0.0939
***p*-value**			<0.01	<0.01	0.22	<0.01	0.46	<0.01	<0.01	0.59	<0.01	<0.01	0.69	0.36	0.08	0.99	<0.01
Class.			I	D	R	I	R	I	I	R	I	D	R	R	I	R	I

Table 64: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the peS2o domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
OpenWebMath	1B	Matched	0.5059	0.5728	0.6168	0.5375	0.5223	0.5981	0.4958	0.5251	0.5343	0.5463	0.5440	0.5446	0.4753	0.4871	0.5361
		Shifted	0.8213	0.2283	0.2819	1.0000	0.3392	1.0000	0.5276	0.3626	0.5267	0.5614	0.9837	0.3975	0.5810	0.4791	0.5779
	7B	Matched	0.5119	0.5835	0.5865	0.5364	0.5167	0.5692	0.4850	0.4730	0.5610	0.5459	0.5277	0.5343	0.4526	0.4514	0.5239
		Shifted	0.8086	0.2216	0.3146	0.9998	0.4781	1.0000	0.3201	0.0000	0.3821	0.4690	0.9500	0.5731	0.7013	0.4365	0.5468
	13B	Matched	0.5212	0.5987	0.6121	0.5460	0.5139	0.6209	0.5039	0.4436	0.5752	0.5549	0.5193	0.5260	0.4922	0.4713	0.5357
		Shifted	0.8242	0.3143	0.3149	0.9874	0.5330	1.0000	0.4841	0.5567	0.3716	0.5568	0.9650	0.5813	0.6993	0.4886	0.6198
	32B	Matched	0.5609	0.6437	0.5901	0.5825	0.6214	0.6364	0.4902	0.4328	0.5867	0.5586	0.5956	0.5457	0.4982	0.4548	0.5570
		Shifted	0.8266	0.3313	0.2911	0.9900	0.6407	1.0000	0.6151	1.0000	0.1926	0.5570	0.9755	0.4435	0.6205	0.5611	0.6461
Avg Δ			+0.2952	−0.3258	−0.3007	+0.4437	−0.0458	+0.3939	−0.0070	+0.0112	−0.1961	−0.0154	+0.4219	−0.0388	+0.1710	+0.0252	+0.0595
***p*-value**			<0.01	<0.01	<0.01	<0.01	0.41	<0.01	0.92	0.96	0.09	0.51	<0.01	0.50	0.02	0.43	0.03
Class.			I	D	D	I	R	I	R	R	D	R	I	R	I	R	I

Table 65: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the OpenWebMath domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	1B	Matched	0.7901	0.8111	0.8081	0.6703	0.5671	0.7479	0.6973	0.8209	0.8520	0.6944	0.6760	0.6412	0.6428	0.6460	0.7189
		Shifted	0.8844	0.9254	0.6556	0.9967	0.3250	1.0000	0.7357	0.5644	0.9615	0.6513	0.6982	0.5150	0.7564	0.5526	0.7302
	7B	Matched	0.8015	0.8406	0.8026	0.7597	0.5806	0.8788	0.6579	0.8386	0.8732	0.6681	0.6857	0.6121	0.6476	0.6575	0.7360
		Shifted	0.8784	0.9216	0.5863	0.9837	0.5868	0.9972	0.7069	0.0000	0.9697	0.5980	0.8128	0.4696	0.8659	0.6055	0.7130
	13B	Matched	0.8075	0.8494	0.7867	0.7759	0.5999	0.8691	0.6358	0.8452	0.8719	0.6743	0.7368	0.6349	0.7022	0.7258	0.7511
		Shifted	0.8799	0.9244	0.5699	0.9674	0.6923	1.0000	0.5886	0.8936	0.9628	0.6505	0.6898	0.5980	0.8354	0.7333	0.7847
	32B	Matched	0.8215	0.8591	0.7997	0.8137	0.6255	0.8824	0.8250	0.8345	0.8670	0.6524	0.6876	0.4149	0.5294	0.7595	0.7409
		Shifted	0.8755	0.9210	0.6041	0.9645	0.5752	0.9976	0.8103	1.0000	0.9510	0.6489	0.7544	0.4101	0.6507	0.7883	0.7823
Avg Δ			+0.0744	+0.0830	−0.1953	+0.2232	−0.0484	+0.1542	+0.0064	−0.2203	+0.0952	−0.0351	+0.0423	−0.0776	+0.1466	−0.0273	+0.0158
***p*-value**			<0.01	<0.01	<0.01	<0.01	0.54	0.02	0.80	0.40	<0.01	0.09	0.33	0.10	<0.01	0.40	0.35
Class.			I	I	D	I	R	I	R	R	I	D	R	R	I	R	R

Table 66: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the StarCoder domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	1B	Matched	0.5732	0.6373	0.5437	0.5661	0.5551	0.5780	0.5138	0.6189	0.6217	0.5549	0.6195	0.4959	0.5001	0.4694	0.5605
		Shifted	0.5723	0.6796	0.5560	1.0000	0.2037	1.0000	0.5535	0.4344	0.7096	0.8166	0.7388	0.4743	0.4946	0.5149	0.6249
	7B	Matched	0.6097	0.6950	0.5850	0.6454	0.6514	0.6512	0.6768	0.6036	0.6590	0.5561	0.5973	0.5144	0.5122	0.4456	0.6002
		Shifted	0.6089	0.7283	0.5954	0.9956	0.6581	1.0000	0.6973	0.0000	0.7963	0.5584	0.6714	0.5148	0.5546	0.4785	0.6327
	13B	Matched	0.6203	0.7213	0.6035	0.6486	0.6758	0.6962	0.7345	0.6536	0.6740	0.5655	0.6767	0.5652	0.5267	0.5130	0.6339
		Shifted	0.6186	0.7519	0.6015	0.9642	0.6769	1.0000	0.7214	0.6881	0.7234	0.8276	0.5729	0.5746	0.5125	0.4981	0.6951
	32B	Matched	0.6459	0.7352	0.5954	0.6729	0.6407	0.7068	0.7739	0.6639	0.7266	0.5767	0.7412	0.7238	0.5113	0.5095	0.6588
		Shifted	0.6511	0.7832	0.5925	0.9752	0.6689	1.0000	0.8572	1.0000	0.7724	0.8289	0.7124	0.7297	0.4592	0.5159	0.7533
Avg Δ			+0.0005	+0.0386	+0.0045	+0.3505	−0.0789	+0.3419	+0.0326	−0.1044	+0.0801	+0.1946	+0.0152	−0.0015	−0.0073	+0.0175	+0.0631
***p*-value**			0.80	<0.01	0.35	<0.01	0.45	<0.01	0.20	0.63	0.03	0.06	0.78	0.85	0.73	0.29	0.02
Class.			R	I	R	I	R	I	R	R	I	I	R	R	R	R	I

Table 67: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the Stack Exchange domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	1B	Matched	0.5893	0.6121	0.4856	0.5674	0.7026	0.5723	0.5218	0.5406	0.5157	0.5199	0.5654	0.4377	0.5718	0.4479	0.5464
		Shifted	0.6510	0.6048	0.4237	0.9884	0.4601	1.0000	0.5285	0.3946	0.5008	0.7074	0.6755	0.4748	0.5551	0.4556	0.6014
	7B	Matched	0.7521	0.7167	0.5694	0.6919	0.7871	0.7095	0.7111	0.7317	0.6342	0.6536	0.5899	0.5362	0.5241	0.4242	0.6451
		Shifted	0.6916	0.6487	0.5662	0.9131	0.7544	0.9919	0.6843	0.0000	0.6104	0.5358	0.5940	0.5011	0.6961	0.4540	0.6173
	13B	Matched	0.6903	0.6197	0.5104	0.6716	0.7177	0.6272	0.6246	0.6371	0.6072	0.6179	0.5535	0.4978	0.5450	0.5033	0.6017
		Shifted	0.6437	0.5886	0.5025	0.8383	0.7277	1.0000	0.6558	0.4062	0.5954	0.6629	0.5262	0.4934	0.6130	0.5088	0.6259
	32B	Matched	0.5865	0.5966	0.5090	0.5802	0.6684	0.5823	0.6155	0.5241	0.5197	0.5169	0.5487	0.5134	0.5425	0.4354	0.5528
		Shifted	0.6204	0.5617	0.4092	0.8878	0.6615	0.9999	0.6978	1.0000	0.5044	0.7030	0.6262	0.5104	0.5429	0.4151	0.6529
Avg Δ			−0.0029	−0.0353	−0.0432	+0.2791	−0.0680	+0.3751	+0.0233	−0.1582	−0.0164	+0.0752	+0.0411	−0.0014	+0.0559	+0.0057	+0.0379
***p*-value**			0.93	0.07	0.16	0.02	0.33	<0.01	0.38	0.57	<0.01	0.38	0.29	0.93	0.28	0.62	0.25
Class.			R	D	R	I	R	I	R	R	D	R	R	R	R	R	R

Table 68: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the SFT domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Stage	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
			PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
DPO	1B	Matched	0.4425	0.5284	0.6443	0.4553	0.4731	0.4535	0.4840	0.5769	0.6041	0.5280	0.5602	0.4592	0.5645	0.4973	0.5194
		Shifted	0.4474	0.5357	0.6304	0.9985	0.2530	1.0000	0.5852	0.6303	0.7095	0.6307	0.5638	0.4482	0.4935	0.4975	0.6017
	7B	Matched	0.4099	0.4342	0.5483	0.3998	0.3857	0.4022	0.5934	0.6276	0.4864	0.4900	0.5653	0.4783	0.5245	0.4519	0.4855
		Shifted	0.4007	0.4924	0.6224	0.8638	0.3860	0.9865	0.6288	0.0000	0.7121	0.5095	0.6096	0.4576	0.5730	0.4498	0.5494
	13B	Matched	0.3853	0.4708	0.6205	0.3858	0.4443	0.3938	0.6637	0.6475	0.5255	0.4816	0.3091	0.4639	0.5992	0.4271	0.4870
		Shifted	0.3861	0.4608	0.6695	0.7009	0.3983	1.0000	0.5068	0.6623	0.6706	0.5923	0.3091	0.4816	0.5268	0.5017	0.5619
	32B	Matched	0.4155	0.4871	0.5893	0.4185	0.4597	0.4025	0.7228	0.6520	0.5890	0.5538	0.2529	0.4946	0.5861	0.4432	0.5048
		Shifted	0.4588	0.5263	0.5957	0.8106	0.4573	0.9888	0.6759	1.0000	0.6963	0.5736	0.5977	0.4815	0.4323	0.4755	0.6265
Avg Δ			+0.0100	+0.0237	+0.0289	+0.4286	−0.0670	+0.5808	−0.0168	−0.0529	+0.1459	+0.0632	+0.0982	−0.0068	−0.0622	+0.0262	+0.0857
***p*-value**			0.45	0.22	0.24	<0.01	0.29	<0.01	0.78	0.81	0.01	0.09	0.32	0.48	0.23	0.24	<0.01
Class.			R	R	R	I	R	I	R	R	I	I	R	R	R	R	I

Table 69: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the DPO domain. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75). The rightmost **Avg** column averages across all methods per row. **Avg** Δ is the mean of (Shifted − Matched) across model sizes. The p-value is from a paired t-test (df $= 3$). **Class.** is the resulting classification: **R** (robust): $p > 0.10$; **I** (inflated): $p \leq 0.10$ and $\overline{\Delta} > 0$; **D** (degraded): $p \leq 0.10$ and $\overline{\Delta} < 0$.

Domain	Subset	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
				PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
StarCoder	Assembly	1B	Matched	0.5833	0.5046	0.3843	0.4136	0.5602	0.3302	0.3611	0.5247	0.5216	0.6451	0.7577	0.4706	0.5077	0.5741	0.5099
			Shifted	0.6960	0.7346	0.5787	1.0000	0.6420	1.0000	0.8642	0.3395	0.7593	0.8920	0.8920	0.6078	0.5355	0.6620	0.7288
		7B	Matched	0.5540	0.5340	0.4969	0.4506	0.6049	0.5062	0.4167	0.5617	0.5679	0.6944	0.5648	0.5458	0.5015	0.5185	0.5370
			Shifted	0.7083	0.7438	0.5432	0.9429	0.6883	0.9660	0.7500	0.0000	0.7716	0.8272	0.8735	0.4902	0.6157	0.4753	0.6711
		13B	Matched	0.5617	0.5062	0.4630	0.4861	0.5170	0.4877	0.4136	0.5247	0.5741	0.7068	0.6944	0.4967	0.5926	0.6667	0.5494
			Shifted	0.7099	0.7253	0.6497	0.8488	0.7284	1.0000	0.7623	0.6975	0.7901	0.8735	0.8565	0.4275	0.5617	0.6327	0.7331
		32B	Matched	0.5710	0.5340	0.4583	0.5355	0.5988	0.5185	0.4537	0.5448	0.6080	0.6914	0.5772	0.4314	0.6358	0.6127	0.5551
			Shifted	0.6728	0.7099	0.5818	0.8333	0.6049	0.9537	0.7407	1.0000	0.7130	0.8765	0.8580	0.5137	0.6296	0.5525	0.7315
	Java	1B	Matched	0.7994	0.8466	0.8320	0.6881	0.5678	0.7787	0.7108	0.8489	0.8751	0.7010	0.5944	0.6562	0.6501	0.6509	0.7286
			Shifted	0.8970	0.9642	0.6536	0.9968	0.3065	1.0000	0.7263	0.5789	0.9811	0.6384	0.5044	0.5072	0.7693	0.5460	0.7193
		7B	Matched	0.8130	0.8781	0.8202	0.7722	0.5855	0.9029	0.6718	0.8547	0.8948	0.6672	0.8066	0.6166	0.6553	0.6654	0.7575
			Shifted	0.8899	0.9538	0.5825	0.9865	0.5843	1.0000	0.7070	0.0000	0.9874	0.5984	0.7521	0.4676	0.8834	0.6127	0.7147
		13B	Matched	0.8213	0.8882	0.8045	0.7872	0.6076	0.8945	0.6476	0.8716	0.8927	0.6740	0.7792	0.6421	0.7082	0.7287	0.7677
			Shifted	0.8928	0.9543	0.5606	0.9723	0.6984	1.0000	0.5770	0.9128	0.9785	0.6380	0.5230	0.6074	0.8549	0.7374	0.7791
		32B	Matched	0.8354	0.8872	0.8148	0.8272	0.6318	0.9080	0.8419	0.8635	0.8877	0.6516	0.7980	0.4122	0.5208	0.7665	0.7605
			Shifted	0.8864	0.9461	0.5993	0.9702	0.5771	1.0000	0.8140	1.0000	0.9692	0.6377	0.6508	0.4028	0.6569	0.7984	0.7792

Table 70: Comparison of AUC between OLMo-Detect (**Matched**) and OLMo-Detect (Shifted) (**Shifted**) on the StarCoder domain by subset. The rightmost **Avg** column averages across all methods per row. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
				PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
Stack Exchange	Math Stack Exchange	1B	Matched	0.6065	0.6531	0.4955	0.5755	0.5033	0.5755	0.6792	0.6184	0.6661	0.5273	0.7388	0.5153	0.6110	0.6498	0.6011
			Shifted	0.5686	0.6792	0.5731	1.0000	0.2882	1.0000	0.6653	0.4000	0.7208	0.7910	0.7388	0.5373	0.5694	0.7249	0.6612
		7B	Matched	0.6249	0.6751	0.5759	0.6800	0.6278	0.6253	0.5200	0.6441	0.7224	0.5380	0.7016	0.4171	0.5090	0.4629	0.5946
			Shifted	0.6449	0.7690	0.5988	1.0000	0.7269	1.0000	0.4808	0.0000	0.8498	0.5176	0.7306	0.4239	0.5735	0.4971	0.6295
		13B	Matched	0.6371	0.7135	0.5690	0.6947	0.6012	0.6963	0.6482	0.6482	0.7510	0.5437	0.6931	0.4960	0.4449	0.5127	0.6178
			Shifted	0.6535	0.8016	0.6245	0.9943	0.7698	1.0000	0.6433	0.7249	0.8082	0.7910	0.6931	0.5357	0.5061	0.5306	0.7198
		32B	Matched	0.6829	0.7494	0.5437	0.7131	0.7008	0.6824	0.7755	0.6735	0.7837	0.5682	0.8518	0.7134	0.3927	0.5784	0.6721
			Shifted	0.7518	0.8759	0.6629	1.0000	0.8420	1.0000	0.8898	1.0000	0.7624	0.7910	0.8518	0.6211	0.3833	0.5486	0.7843
	Stack Overflow	1B	Matched	0.5768	0.6356	0.5480	0.5647	0.5605	0.5806	0.5045	0.6219	0.6192	0.5570	0.5003	0.4949	0.4895	0.4501	0.5503
			Shifted	0.5781	0.6797	0.5561	1.0000	0.1898	1.0000	0.5551	0.4373	0.7090	0.8191	0.4616	0.4687	0.4872	0.4927	0.6025
		7B	Matched	0.6151	0.6980	0.5860	0.6462	0.6623	0.6595	0.6956	0.6064	0.6546	0.5582	0.4931	0.5214	0.5128	0.4425	0.5965
			Shifted	0.6132	0.7246	0.5952	0.9956	0.6612	1.0000	0.7226	0.0000	0.7927	0.5636	0.6122	0.5241	0.5540	0.4758	0.6311
		13B	Matched	0.6277	0.7225	0.6067	0.6534	0.6844	0.7045	0.7476	0.6590	0.6692	0.5674	0.6603	0.5726	0.5335	0.5133	0.6373
			Shifted	0.6254	0.7467	0.5997	0.9687	0.6718	1.0000	0.7354	0.6866	0.7184	0.8310	0.4526	0.5761	0.5119	0.4939	0.6870
		32B	Matched	0.6583	0.7356	0.5996	0.6862	0.6341	0.7177	0.7905	0.6752	0.7236	0.5783	0.6305	0.7255	0.5211	0.5027	0.6556
			Shifted	0.6611	0.7771	0.5869	0.9812	0.6530	1.0000	0.8726	1.0000	0.7748	0.8325	0.5730	0.7408	0.4667	0.5105	0.7450

Table 71: Comparison of AUC between OLMO-DETECT (**Matched**) and OLMO-DETECT (SHIFTED) (**Shifted**) on the Stack Exchange domain by subset. The rightmost **Avg** column averages across all methods per row. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).

Domain	Subset	Size	Setting	Likelihood-based											Pert.	Out.-Dist.	Prompt	Avg
				PPL	Zlib	Lower	MinK	MinK++	DCPDD	RECALL	CAMIA	PAC	Neigh.	EMMIA	DCQ	CDD	Guided	
SFT	Aya	1B	Matched	0.5436	0.5506	0.4896	0.5018	0.6373	0.5031	0.4558	0.5861	0.4859	0.4934	0.5811	0.4183	0.6015	0.4148	0.5188
			Shifted	0.6315	0.5327	0.4017	0.9876	0.4307	1.0000	0.4929	0.4631	0.4032	0.8260	0.7061	0.4821	0.5787	0.4116	0.5963
		7B	Matched	0.7345	0.6066	0.5135	0.6517	0.7674	0.6481	0.7063	0.7277	0.5803	0.6689	0.5872	0.5785	0.5338	0.4366	0.6244
			Shifted	0.7075	0.5500	0.5145	0.9265	0.7615	1.0000	0.6612	0.0000	0.4664	0.5658	0.5900	0.5473	0.7517	0.4441	0.6062
		13B	Matched	0.6773	0.5107	0.4521	0.6447	0.6504	0.5473	0.6065	0.6166	0.5348	0.6002	0.5094	0.5028	0.5551	0.4461	0.5610
			Shifted	0.6602	0.5028	0.4578	0.8435	0.7155	1.0000	0.6971	0.4193	0.4552	0.8045	0.5242	0.4932	0.6709	0.4748	0.6228
		32B	Matched	0.5512	0.5406	0.5175	0.5367	0.6315	0.5443	0.6447	0.4834	0.4882	0.4833	0.4351	0.5223	0.5423	0.3777	0.5213
			Shifted	0.6148	0.4894	0.3956	0.8859	0.6374	1.0000	0.7661	1.0000	0.4054	0.8154	0.6133	0.5006	0.5737	0.3571	0.6468
	WildChat	1B	Matched	0.7179	0.7725	0.5031	0.7385	0.8694	0.7415	0.6880	0.4622	0.6289	0.5828	0.5498	0.4496	0.5062	0.4936	0.6217
			Shifted	0.6682	0.8027	0.5360	1.0000	0.5428	1.0000	0.6287	0.2451	0.7681	0.5266	0.6450	0.5158	0.5018	0.5412	0.6373
		7B	Matched	0.7894	0.8799	0.6899	0.7485	0.8511	0.7984	0.7181	0.7727	0.7316	0.6434	0.5927	0.4670	0.5129	0.4127	0.6863
			Shifted	0.6827	0.7893	0.6966	0.9040	0.7701	0.9803	0.7179	0.0000	0.7984	0.5016	0.5980	0.4532	0.6323	0.4719	0.6426
		13B	Matched	0.7223	0.8147	0.6312	0.7076	0.8272	0.7328	0.6455	0.7101	0.7082	0.6323	0.5976	0.5006	0.5302	0.5849	0.6675
			Shifted	0.6156	0.7454	0.6167	0.8421	0.7518	1.0000	0.6097	0.3708	0.7739	0.5448	0.5281	0.5015	0.5424	0.5645	0.6434
		32B	Matched	0.7261	0.7629	0.5139	0.7643	0.7941	0.7075	0.6445	0.6523	0.6028	0.5918	0.6624	0.4685	0.5431	0.5466	0.6415
			Shifted	0.6289	0.7861	0.4930	0.9667	0.7363	1.0000	0.6621	1.0000	0.7415	0.5371	0.6392	0.5515	0.4756	0.5325	0.6965

Table 72: Comparison of AUC between OLMO-DETECT (**Matched**) and OLMO-DETECT (SHIFTED) (**Shifted**) on the SFT domain by subset. The rightmost **Avg** column averages across all methods per row. Cells are colored by magnitude: weak (< 0.60), moderate (0.60–0.75), and strong (≥ 0.75).