Where does output diversity collapse in post-training?

Constantinos Karouzos **Xingwei Tan** **Nikolaos Aletras**

School of Computer Science
University of Sheffield, UK
{kkarouzos1, xingwei.tan, n.aletras}@sheffield.ac.uk

arXiv:2604.16027v1 [cs.CL] 17 Apr 2026

Abstract

Post-trained language models produce less varied outputs than their base counterparts. This output diversity collapse undermines inference-time scaling methods that rely on varied samples, and risks homogenizing model outputs on creative and value-laden tasks. Prior work attributes collapse to specific post-training methods, without separating the role of training data composition from the method, or the generation format from the model weights. We trace output diversity through three parallel post-training lineages of Olmo 3, Think (chain-of-thought distillation), Instruct (broad multi-source data), and RL-Zero, across 15 tasks and four text diversity metrics. We find that the location of collapse co-varies with data composition: the Think lineage loses most semantic diversity at supervised fine-tuning, and the effect of DPO is larger in Instruct than in Think. Suppressing chain-of-thought reasoning at inference in Think models drops accuracy on hard tasks, yet leaves answer-level diversity unchanged, showing that the collapse is embedded in the model weights by training data, not imposed by the generation format. Decomposing diversity loss on six verifiable tasks into a quality-control component (removal of incorrect outputs) and a residual component (genuine narrowing among correct outputs) reveals that the split is task-dependent, and Think models retain more correct-answer diversity than Instruct despite collapsing more in aggregate. Our results indicate that diversity collapse is determined during training by data composition and cannot be addressed at inference time alone.[1]

1 Introduction

Large language models (LLMs) rely on post-training to improve helpfulness, safety, and instruction compliance. Post-training combines supervised fine-tuning (SFT; Ouyang et al., 2022) on curated demonstrations, and direct preference optimization (DPO; Rafailov et al., 2023) or reinforcement learning from human feedback (RLHF). However, this results in output diversity collapse, i.e., models produce more uniform outputs than their base counterparts across summarization (Kirk et al., 2024b), reasoning (Dang et al., 2025), and open-ended generation (Jiang et al., 2025). Diversity collapse limits self-consistency (Wang et al., 2023), pass@k sampling (Chen et al., 2021), and test-time compute scaling (Snell et al., 2025). Kamigaito et al. (2025) show diversity is the mechanism underlying inference scaling laws. The algorithmic causes are well-understood (Wang et al., 2024; Ma et al., 2025a; GX-Chen et al., 2026), yet diversity collapses across task types. This leads LLMs to produce less diverse outputs than a basic web search (Wright et al., 2026), co-writing with LLMs reduces content diversity (Padmakumar & He, 2024), and single-reward RLHF can amplify majority preferences to near-total dominance (Chakraborty et al., 2024).

Yet, prior work attributes collapse to specific algorithms. DPO in narrative generation (Peeperkorn et al., 2025), the reward step in creative tasks (O'Mahony et al., 2024), and SFT in reasoning (Dang et al., 2025), without investigating the effect of *data* compositions. Ma

[1]Code: https://github.com/ckarouzos/where-diversity-collapses/

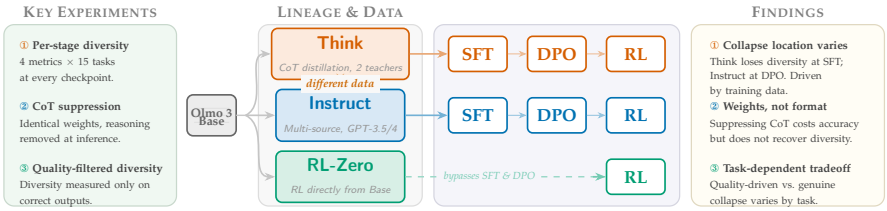

Figure 1: Study design. We trace output diversity through three parallel post-training lineages of Olmo 3, to identify where, why, and how much diversity is lost.

et al. (2025b) suppress chain-of-thought (CoT; Wei et al., 2022) at inference but measure only accuracy, not diversity. No existing study isolates the role of the training *method* from the training *data*, or the generation *format* from the model weights.

Two questions remain open: (1) *does the diversity collapse co-vary with the post-training method or with the post-training data composition*, and (2) *does the CoT format itself constrain diversity at inference, or is the collapse embedded in the model weights?*

We answer these questions through a controlled experimental setting (Figure 1). We monitor the output diversity of the open weight and data Olmo 3 model family (Olmo et al., 2025), which releases checkpoints of all post-training stages across three parallel lines. **Think** and **Instruct** variants share the same post-training recipe (SFT→DPO→RL) but differ in data, while **RL-Zero** bypasses SFT and DPO entirely. Evaluating 13 models across 15 tasks with four diversity metrics, we show that the same post-training method produces different diversity outcomes depending on the upstream data composition, and that each stage plays a distinct role. Our contributions:

- We compare Think vs. Instruct lineages, showing that collapse location depends on data: narrow CoT distillation for Think models is associated with a larger drop at SFT, while the DPO drop is larger in Instruct models (§4.1);
- We evaluate Think models with CoT suppressed at inference and find no diversity recovery on any task–stage combinations, while quality drops. Diversity collapse resides in the model weights, not in the CoT generation format (§4.2);
- We decompose diversity reduction into a quality-control component (removal of incorrect outputs) and a residual component (genuine narrowing among correct outputs), showing the split is task-dependent (§4.3).

2 Related work

The reliability–diversity tradeoff in post-training. Jiang et al. (2025) show that aligned models exhibit high output homogeneity across a wide range of model families and scales. Kirk et al. (2024b) find that RLHF reduces both per input and across input diversity. Human co-writing with aligned models reduces content diversity (Padmakumar & He, 2024), and users brainstorming with ChatGPT produce less semantically distinct ideas (Anderson et al., 2024). In reasoning, SFT improves pass@1 but degrades pass@k (Dang et al., 2025); base models outperform RLVR-trained models at large sample budgets (Yue et al., 2025), and base models produce more diverse outputs (West & Potts, 2025). Peeperkorn et al. (2025) identified DPO as the steepest drop. Karouzos et al. (2026) show that under domain shift the adaptation strategy dominates the alignment objective. Current methods cannot selectively preserve diversity where it is beneficial (Jain et al., 2025). Quality-adjusted diversity shows that preference-tuned models retain higher diversity among high-quality outputs (Shypula et al., 2025), and multi-dimensional linguistic benchmarks find that larger models are often less diverse than smaller ones (Guo et al., 2025b). Automatic diversity metrics lag behind human judgments (Tevet & Berant, 2021), and sampling temperature cannot recover training-induced loss (Verine et al., 2025).

Mechanisms and mitigations. DPO's gradient imbalance suppresses dispreferred responses (Ma et al., 2025a), and likelihood displacement shifts probability to unintended outputs (Razin et al., 2025). KL-regularized RL specifies unimodal targets by construction (GX-Chen et al., 2026), preference collapse arises from KL amplification (Xiao et al.,

2024), and chat templates induce diversity collapse (Yun et al., 2025). Training on recursively generated synthetic data causes progressive tail disappearance (Shumailov et al., 2024). Proposed mitigations include forward-KL optimization (Wang et al., 2024), entropy-constrained RL (Pan et al., 2026), decoupled regularization (Slocum et al., 2025), game-theoretic SFT (Li et al., 2025c), diversity-aware preference optimization (Li et al., 2025a; Lanchantin et al., 2025), and conformative decoding (Peeperkorn et al., 2025). A single reward function is insufficient to represent diverse human preferences (Chakraborty et al., 2024).

3 Experimental setup

3.1 Models and training lineages

We study 13 Olmo 3 checkpoints at the 7B scale. Post-training applies up to three stages, SFT, DPO, and RL, starting from the same base model.

Base (1 model). The base model is pretrained on Dolma 3 Mix (6T tokens), midtrained on Dolmino Mix (100B tokens), and context-extended to 65K tokens.

Think (3 models: Think-SFT, Think-DPO, Think). SFT trains on ∼2.3M synthetic CoT (Wei et al., 2022) reasoning traces using (prompt, completion) pairs from two teachers: QwQ-32B (Team, 2024) and DeepSeek-R1 (Guo et al., 2025a). DPO uses ∼200K Delta Learning (Geng et al., 2025) pairs. The RL stage uses a variation of GRPO (Shao et al., 2024) with verifiable rewards and no KL penalty, and trains on ∼105K prompts, to produce Think.

Think-not-thinking. To isolate the contribution of the CoT generation format from the learned weights, we additionally evaluate all three Think checkpoints with CoT suppressed by prefilling an empty `<think>\n</think>\n` block, forcing direct answers.

Instruct (3 models: Instruct-SFT, Instruct-DPO, Instruct). SFT *initializes from* Think-SFT, then trains on ∼2.2M examples that include function-calling, strip reasoning traces, and draw from multiple sources (GPT-3.5, GPT-4, GPT-4.1; OpenAI et al., 2024) rather than two teachers. DPO (∼260K pairs) uses the same pool of prompts as Think-DPO but with the thinking mode disabled, adding multi-turn and GPT-judged preference pairs. The same RL stage as Think produces the final Instruct model.

RL-Zero (6 models). Applies RL training directly to Base, bypassing SFT and DPO. Four Olmo 3 variants target different reward domains: RL-Zero-Math, RL-Zero-Code, RL-Zero-IF, and RL-Zero-General (∼105K prompts each). Two additional Olmo 3.1 variants (RL-Zero-Math$^{3.1}$, RL-Zero-Code$^{3.1}$) are trained for more steps.

3.2 Tasks and Data

Summarization. TL;DR (Völske et al., 2017), CNN/DailyMail (Nallapati et al., 2016), and XSum (Narayan et al., 2018). Bounded output length controls for length confounds, and multiple valid summaries provide a clear diversity signal.

Code. HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and CRUXEval (Gu et al., 2024). Outputs can be syntactically different but functionally identical, and RL directly optimizes code tasks.

Reasoning. GSM8K (Cobbe et al., 2021), MATH-Algebra, MATH-Geometry (Hendrycks et al., 2021), and TruthfulQA (Lin et al., 2022), the primary Think and RL-Zero training domain. Diversity here measures variation in solution *strategy* with answers held constant.

Instruction following. Alpaca (Taori et al., 2023), open-ended, and IFEval (Zhou et al., 2023), with verifiable format constraints.

Creative writing. WritingPrompts (Fan et al., 2018), where diversity is intrinsically desirable.

Value pluralism. PRISM (Kirk et al., 2024a) and WildBench (Lin et al., 2025), which test whether alignment imposes a single perspective on contested topics.

We measure training–evaluation overlap using C_{13} 13-gram matching (Lambert et al., 2025) between the four Dolci post-training datasets and all fifteen evaluation tasks (Appendix J). Nine datasets show negligible overlap ($\leq 2\%$). HumanEval, CRUXEval, IFEval, MATH-Algebra, MATH-Geometry, and WildBench show elevated overlap (7–30%), traceable to shared upstream data. While we flag these benchmarks, our findings on contaminated tasks are consistent with the patterns on the clean tasks.

3.3 Metrics

We measure diversity along four complementary axes (detailed definitions in Appendix B). **EAD** (Liu et al., 2022) counts unique n-grams normalized against the expected count under a uniform draw (averaged over $n \in \{1, \ldots, 5\}$), capturing *lexical* diversity. **SBERT** computes mean pairwise cosine distance of sentence embeddings (`all-mpnet-base-v2`; Reimers & Gurevych, 2019), capturing *semantic* diversity (0 = collapse, 1 = dissimilar). For code tasks we additionally report *semantic* diversity with UniXcoder (Guo et al., 2022) embeddings (Appendix F). **NLI** scores output pairs with an NLI classifier (`roberta-large-mnli`; Liu et al., 2019), following Stasaski & Hearst (2022), capturing *logical* diversity; code tasks are excluded. **Vendi Score** (Friedman & Dieng, 2023) measures the effective number of dissimilar outputs via eigenvalue entropy of the SBERT similarity kernel (VS=1: identical, VS=K: orthogonal). For code-generation tasks we also report **AST subtree diversity**, the mean pairwise Jaccard distance on AST subtree multisets (Shypula et al., 2025), on correct outputs only (Appendix F).

Quality. For the six tasks with verifiable answers (GSM8K, MATH-Algebra, MATH-Geometry, HumanEval, MBPP, IFEval), we report: accuracy@1 (greedy decoding), majority vote@16 (most frequent answer among K=16 samples), and pass@16 (at least one correct among K). For code tasks we use the unbiased pass@k estimator. For IFEval we report strict and loose constraint satisfaction. For the eight tasks without verifiable answers we evaluate quality using LLM-as-judge (`gpt-4.1-mini`) with established protocols (Appendix D).

Quality-filtered diversity. We decompose diversity into a quality-control component (removal of incorrect outputs) and a residual component (genuine narrowing among correct outputs). D_a (SBERT on all K outputs) and D_c (SBERT on the $K_c \geq 2$ correct outputs). The gap $D_a - D_c$ reflects diversity from error variety; D_c captures genuine narrowing among correct solutions. We report analogous Vendi scores V_a and V_c.

For each model–task pair, we generate K=16 outputs per prompt at T=0.6, top-p=0.95. Base recommends T=1.0; we use matched settings for controlled comparison (Appendix H). For all Think-lineage models, we strip `<think>...</think>` reasoning traces before computing any metric, so that all diversity and quality scores reflect the *final answer* only. Implementation details are in Appendix A.

4 Results

We present results around three questions. First, *where* does diversity collapse along each lineage (§4.1; Figure 2, Table 1)? Second, does the CoT generation format itself constrain diversity (§4.2; Figures 4–5)? Third, how much of the observed collapse is attributable to quality control (§4.3; Figures 6–8)?

4.1 Lineage-dependent diversity collapse

SFT asymmetry. Think and Instruct share the same three-stage post-training, yet collapse at different stages. Think-SFT loses 62% (Table 1) of Base diversity on average, 24% more than Instruct-SFT (38%), uniformly across all 15 tasks, consistent with *completion homogeneity* from two teachers rather than prompt overlap. This challenges findings of minimal SFT impact on diversity (Guo et al., 2025b) and suggests that the effect depends on the breadth of the SFT data. Collapse magnitude also scales with task difficulty (Figure 2). Think-SFT retains only 36% of Base diversity on GSM8K (92% accuracy) but 54% on MATH-Geometry (50% accuracy). Easier tasks with a dominant solution strategy collapse the most. Instruct-SFT,

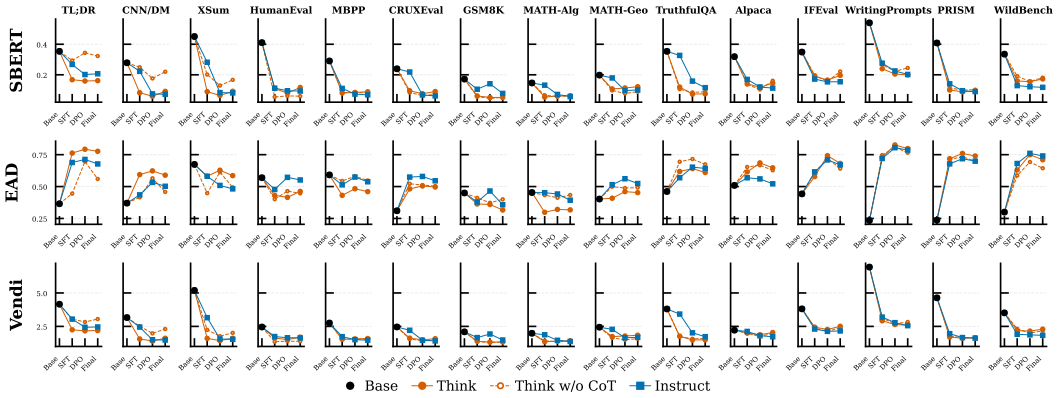

Figure 2: SBERT, EAD, and Vendi Score across post-training stages. Think (orange) collapses at SFT; Instruct (blue) at DPO. Think w/o CoT (hollow) tracks Think.

despite initializing from the already-collapsed Think-SFT, recovers a median 40% of the lost diversity, likely due to its multi-source data. As Instruct-SFT initializes from Think-SFT, this recovery also reflects the dynamics of retraining a collapsed model.

DPO asymmetry. DPO erases more diversity in Instruct than in Think, as Think has already collapsed at SFT, leaving little for DPO to remove. The effect is largest on summarization and code-reasoning tasks, where Instruct-SFT had preserved substantial diversity. On three math/code tasks, Think-DPO actually *increases* diversity slightly, and Instruct-DPO does the same on GSM8K, suggesting that DPO can partially correct a collapsed SFT distribution.

	SFT	DPO	RL	Retained
Think	−62	−4	+4	38%
Instruct	−38	−23	−5	34%
RL-Zero	(single)			93%

Table 1: Stage-wise SBERT loss (% of Base, 15-task average).

RL reversal. Think's RL stage increases semantic diversity on most tasks, primarily code and summarization. The recovery is modest (roughly 5% of total diversity lost) but directionally consistent. Both lineages use the same RLVR method, so the asymmetry likely reflects the input state: Think enters RL already at its diversity floor, leaving room for exploration, while Instruct enters with residual diversity that RL continues to compress. On GSM8K, Instruct RL erases 37% of Base diversity, the largest single-stage loss outside SFT, as the verifiable reward concentrates probability on the dominant correct strategy. The RLVR stage also produces lexically *more uniform* outputs (EAD decreases on nearly all tasks), suggesting it standardizes surface form while broadening semantic content.

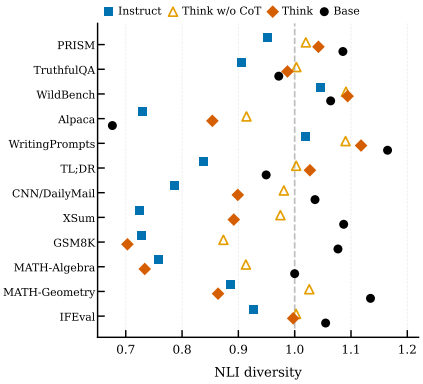

Figure 3: NLI diversity.

Convergence. RL-Zero bypasses both bottlenecks (Figure 2), retaining $\geq$ 71% of Base diversity (median 94%). Both supervised lineages converge to similar final diversity floors (with Think slightly higher on 11/15 tasks), despite different trajectories: data composition co-varies with *when* and *how sharply* diversity is lost. Table 1 summarizes the stage-wise attribution. Full per-task breakdowns are in Appendix I.

The collapse is semantic, not lexical (Figure 2). Per input SBERT drops from 0.32 (Base) to 0.12 (Think) and 0.11 (Instruct), and the Vendi Score drops from ∼3.4 effective modes to ∼1.8 (final), with near-total collapse on math (GSM8K: 1.3 modes, MATH-Algebra: 1.4), 16 samples carry essentially no more semantic diversity than one. EAD (Figure 2) remains stable or *increases*, even as semantic diversity drops. Aligned models use varied vocabulary and phrasing to express semantically identical content. Think's EAD on WritingPrompts rises from 0.23 to 0.80, while SBERT falls from 0.54 to 0.20, a pattern replicated across open-ended tasks. For natural language tasks, NLI diversity (Figure 3) drops on most tasks,

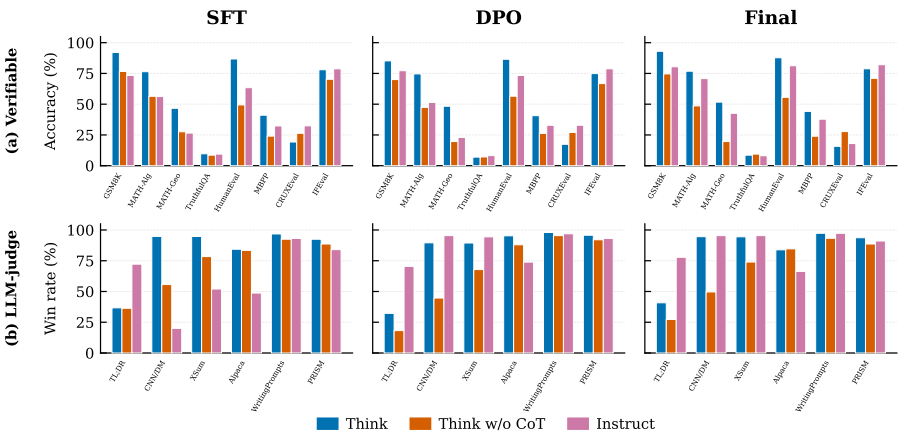

Figure 4: Quality of generations for Think, Think-not-thinking, and Instruct, across stages. **Top**: accuracy on eight verifiable tasks. **Bottom**: LLM-judge win rates on six tasks.

though the gap varies. Post-trained models still make logically distinct claims. The gap is largest for Think models, where CoT reasoning preserves logical structure even as the surface distribution narrows.

Value-pluralism tasks suffer the steepest Think collapse (PRISM −78%, TruthfulQA −79%), as narrow two-teacher distillation cannot represent the range of perspectives these tasks require. On PRISM, Think's NLI (Figure 3) scores remain above 1.0 (net contradictions), meaning the model still samples contradictions despite converged phrasing, though we cannot determine whether this is genuine stance plurality or internal incoherence. Instruct drops NLI below 1.0, indicating homogenization of both form and stance (Figure 3). Think's NLI remains above the contradiction threshold on value-pluralism and creative tasks where Instruct's drops below. Creative writing (WritingPrompts) shows the highest Base diversity (6.9 Vendi modes) and the sharpest quality–diversity tension. Think and Instruct both collapse to ∼0.20 SBERT and ∼2.6 modes (−63%), yet achieve >97% pairwise win rate against Base, producing better stories at the cost of formulaic variation. RL-Zero retains ∼100% of Base diversity, but wins only ∼50%, consistent with the absence of a creative-writing reward signal. NLI diversity remains above 1.0 for all models on WritingPrompts (Think 1.12, Instruct 1.02, RL-Zero 1.15), meaning post-trained models still produce logically distinct narratives despite semantic convergence. Full per-task breakdowns are in Appendix C.

4.2 Think-not-thinking: CoT as reliability, not diversity

Think and Instruct differ in both training data *and* generation format. Think generates CoT reasoning traces before answering, while Instruct answers directly. To isolate the format's contribution, we evaluate all three Think models with CoT suppressed, we refer to these models as *Think-not-thinking*. This is an out-of-distribution intervention, so we interpret the results as testing whether format removal recover diversity. Across tasks (Figure 2), removing CoT **does not recover diversity**. Think-not-thinking SBERT diversity matches Think, and Instruct shows similarly collapsed diversity. This holds at every stage (SFT, DPO, RLVR) and across every task category. IFEval shows a small increase (+0.025 SBERT), but this is modest relative to the Base-to-Think gap (−0.153).

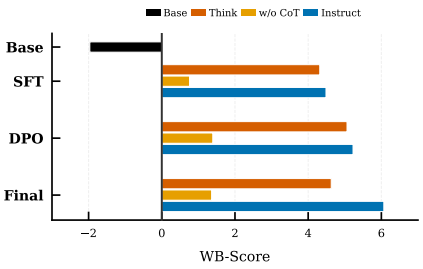

Figure 5: WildBench Score.

CoT suppression *does* affect accuracy (Figure 4), with harder tasks losing more: IFEval −8%, GSM8K −18%, MBPP −20%, MATH-Algebra −28%, HumanEval −32%, MATH-Geometry −32%. The quality cost is task-dependent (Figure 4), CoT suppression is negligible for open-ended generation (no change for Alpaca, WritingPrompts −4%) but severe for summarization (CNN/DM −48%) and complex helpfulness (WildBench Score 4.6 → 1.4, Figure 5). In no case does suppression recover diversity. CoT improves reliability by helping

the model execute its learned strategy, especially on hard problems, without broadening the answer-level diversity distribution. The output distribution is equally collapsed whether the model reasons explicitly or answers directly. One exception is WritingPrompts, where removing CoTs slightly *increases* SBERT diversity (+0.046), suggesting that CoT imposes implicit narrative templates that constrain story generation. NLI diversity reveals a subtler pattern on math tasks: Think-not-thinking produces *higher* NLI scores than Think (GSM8K: 0.87 vs. 0.70; MATH-Algebra: 0.91 vs. 0.73), despite identical SBERT. Without CoT, final answers are semantically collapsed but logically less entailing. The model generates diverse wrong answers rather than diverse correct strategies, consistent with the accuracy drops.

Diversity collapse resides in the learned distribution, not the output format. Narrow two-teacher SFT data reshapes model outputs, and this effect is not reversed by suppressing CoT at inference. This aligns with findings that CoT in post-trained models can function as post-hoc rationalization (Lewis-Lim et al., 2025) and that CoT can be applied selectively (Sprague et al., 2025). The model has already converged on its answer distribution during training. The Think vs Instruct comparison (§4.1) is, therefore, not confounded by the generation format. The diversity difference between lineages reflects data composition. *Practitioners cannot recover diversity by switching Think models to direct-answer mode, the cost is paid at training time*. We note that we measure final-answer diversity, not reasoning-path diversity.

4.3 Quality-filtered diversity decomposition

The aggregate diversity reductions combine two effects, elimination of incorrect outputs and genuine narrowing of the correct-answer distribution (Figure 6). We decompose these using D_a, D_c, V_a and V_c on six verifiable tasks (GSM8K, MATH-Algebra, MATH-Geometry, HumanEval, MBPP, IFEval). All models achieve 94–97% pass@16 on GSM8K, the underlying capability is broadly present. RL-Zero variants also reach 94–97% pass@16 on GSM8K despite 49–61% accuracy@1, confirming the gap is in reliability, not capability. The difference lies in per-attempt reliability (Think 93% vs. Base 56%), not in whether the knowledge exists.

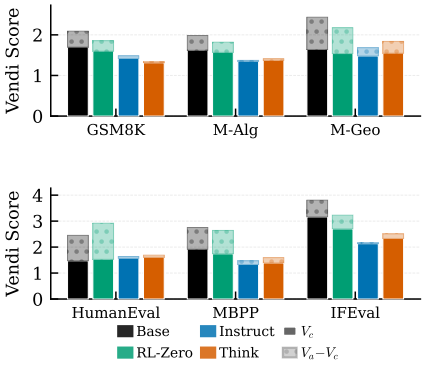

Figure 6: Quality filtered Vendi Score on six verifiable tasks.

The proportion of collapse attributable to quality control varies by task (Figure 6; Appendix E): on IFEval, 83.4% of the D_a drop persists in D_c (genuine narrowing), while on MBPP 38% is genuine and on HumanEval less than 10%. Math reasoning falls between (57–64% genuine). Code-specific metrics sharpen this picture: among correct HumanEval outputs, Think produces structurally homogeneous solutions (AST Jaccard =0.53, UniXcoder D_c=0.13) while Base/RL-Zero's correct outputs are structurally diverse (AST Jaccard =0.89 on MBPP; Figure 7). This resolves the tension between *diversity collapse is harmful* and *it is just quality control* (Lake et al., 2025): both are right, in task-dependent proportions.

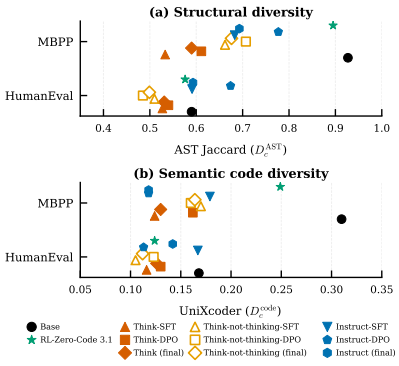

Figure 7: Code diversity on correct outputs: AST subtree Jaccard (structural) and UniXcoder (semantic) for HumanEval and MBPP.

Even among correct outputs, a narrowing persists: Base maintains 1.7 effective Vendi modes among its ∼8.5/16 correct answers, while both Think and Instruct converge to 1.3–1.6 modes among their correct answers (∼15/16 for GSM8K), while IFEval is higher at 2.1–2.3. In absolute terms, all post-trained models produce near-homogeneous correct outputs, which limits the effectiveness of majority voting (Wang et al., 2023): Think gains just +0.4% on GSM8K (16 near-identical correct answers provide no independent signal), while Base gains +24% and RL-Zero +22–26%. Correct-answer diversity determines how

much models benefit from repeated sampling (Snell et al., 2025). On MATH-Algebra, Think-not-thinking and RL-Zero-Math both achieve ∼49% accuracy, but RL-Zero-Math has twice the correct-answer diversity and gains +15% from majority voting compared to +7% for Think-not-thinking. The pattern holds across math tasks (Figure 8): at matched accuracy, models with more diverse correct outputs consistently extract more benefit from sampling.

On HumanEval, Instruct surpasses Think at pass@16 (98.2 vs. 95.7) despite trailing at pass@1 (81.2 vs. 87.7). The collapsed output distribution means additional samples yield identical solutions. On TruthfulQA, the effect is reversed, majority-voting actually *hurts* all models (majority vote@16 < accuracy@1), because the model converges confidently onto the misconception the question was designed to test. When the dominant mode is wrong, diversity collapse amplifies the error. Figure 8 visualizes this pattern, high-accuracy models cluster near zero MV gain, while lower-accuracy models with diverse correct outputs benefit substantially. Full quality results are in Appendix D; quality-filtered results in Appendix E.

4.4 Cross-cutting patterns

The ordering (Base > RL-Zero > Final) holds on average across all 15 tasks, though individual RL-Zero variants exceed Base on tasks aligned with their reward signal (e.g., RL-Zero-IF on IFEval, RL-Zero-Code[3.1] on HumanEval). A model that is low-diversity on one task tends to be low-diversity on all tasks. Output length does not explain diversity ordering (Appendix G).

LLM-as-a-judge evaluation (Figure 4) confirms post-training improves quality across all non-verified task categories. CNN/DM and XSum win rates increase from 26–48% (Base) to 83–95% (Think, Instruct), open-ended pairwise win rates exceed 80% for Think on Alpaca and for both Think and Instruct on PRISM. WildBench scores rise from −2.0 (Base) to 6.1 (Instruct). RL-Zero models are tied with Base on WritingPrompts (50% win rate), consistent with the absence of creative-writing reward signals. Diversity reductions coexist with clear quality gains.

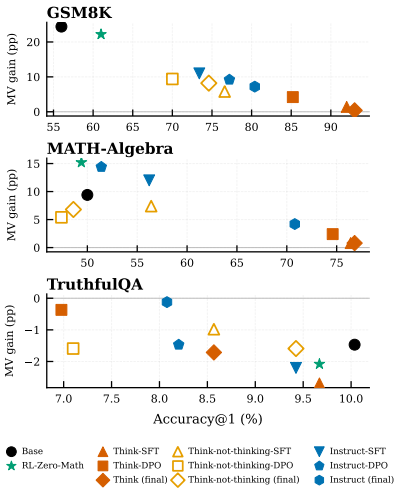

Figure 8: Accuracy@1 vs. majority-voting gain.

Among RL-Zero variants, the reward signal type predicts diversity preservation. RL-Zero-IF (instruction-following rewards) retains 99% of Base diversity on average, while RL-Zero-Code retains only 88%. On code tasks specifically, RL-Zero-Code retains *less* diversity (90%) than RL-Zero-General (100%). Pass/fail execution rewards narrow the solution space more aggressively than general rewards. Mathematical reasoning rewards, which admit diverse solution paths, fall between these extremes. This order (format rewards > math rewards > code rewards) shows that the reward specificity predicts diversity reduction. However, RL-Zero's diversity advantage comes at a steep quality cost, the RL-Zero range is 49.8-61.0% on GSM8K (vs. 93% Think, 80% Instruct) and 49% on IFEval (vs. 79% Think).

5 Discussion

Data composition co-varies with the trajectory, not the floor. Think and Instruct share the same three-stage training yet collapse at different stages. The DPO asymmetry (§4.1) reflects the upstream SFT state more than DPO data differences. Think collapses uniformly across all tasks at SFT, leaving DPO little to remove, while Instruct enters DPO with residual spread that is aggressively narrowed. Despite these different paths, both lineages converge to 1.3–1.6 Vendi modes among correct answers on most verifiable tasks and ∼2 modes overall, with IFEval as an outlier at 2.1–2.3. *Data composition* determines *when* and *how sharply* models reach the diversity floor, but not the floor itself. This distinction matters practically, data-level interventions (more teachers, broader sources) can slow the descent

but may not raise the final diversity level. Algorithmic changes, switching from reverse to forward KL (Wang et al., 2024), adding entropy constraints (Pan et al., 2026), or removing KL penalties entirely (as in RL-Zero), appear necessary to shift the floor. For SFT data, this suggests that the number of distinct completion sources matters. *Practitioners should avoid single-teacher or dual-teacher distillation when output diversity is valued, and instead draw from multiple models with diverse training*.

Mechanistic interpretation. SFT via cross-entropy loss on narrow data performs maximum-likelihood estimation on a low-entropy target distribution. As two teachers from related training lineages produce completions occupying a restricted region of the output space, the model reproduces this narrow mixture. DPO's reverse-KL objective is mode-seeking by construction, its gradient is proportional to the implicit reward gap between chosen and rejected outputs. When the model is already collapsed (Think post-SFT), chosen and rejected responses are both near the mode, yielding small gradients and minimal further compression. When the model retains spread (Instruct post-SFT), DPO aggressively downweights the tails. GRPO *without KL regularization* frees the policy to rediscover modes that SFT and DPO suppressed, provided they receive a positive reward signal.

Task-dependent patterns: where diversity loss matters most. On *math and reasoning tasks* a significant part of diversity reduction reflects removal of incorrect solution paths, as the narrowing among correct outputs is modest. On *code* tasks, less collapse is genuine narrowing, but it still limits pass@k scaling. *Summarization* shows the largest semantic diversity loss, but this is the cost for large quality gains. *Creative writing and value-pluralism* are the tasks where the observed diversity loss risks imposing a single perspective. The pattern that emerges is a spectrum, from tasks where collapse is largely helpful (code correctness filtering) to tasks where it is actively harmful (value-laden open-ended generation). *Practitioners should assess diversity impact relative to their task characteristics, when selecting post-trained models or applying uniform post-training recipes*.

From distributional to representational diversity. We capture *distributional* diversity, i.e. statistical spread along lexical, semantic, and logical axes. This is not a sufficient condition for *representational* diversity, the presence of outputs reflecting different perspectives or stances. We detect when a model's output distribution narrows but cannot determine which perspectives are lost. The distinction matters most on value-pluralism tasks. Narrow training data does not just reduce variation, it risks imposing a single perspective on questions where legitimate disagreement exists. A model could maintain high distributional diversity while eliminating viewpoints, or conversely appear collapsed while preserving the stances that matter most. Targeted probes for representational diversity across demographic and cultural dimensions are needed to close this gap.

6 Conclusion

We traced output diversity through three parallel post-training lineages of Olmo 3, showing that diversity collapse is shaped by training data composition, not the post-training method alone. The same three-stage recipe (SFT→DPO→RL) produces different collapse trajectories depending on the upstream data: narrow two-teacher distillation drives a steep SFT cliff, while broader multi-source data shifts the sharpest drop to DPO. Suppressing the CoT generation format at inference costs accuracy, but does not recover diversity, confirming that the collapse resides in the learned weights. Decomposing the diversity loss into quality-control and residual components reveals a task-dependent split. On some tasks nearly all narrowing reflects the removal of errors, on others most of it is genuine homogenization among correct outputs. This directly affects inference scaling and majority voting boosts. For practitioners, our results point to two actionable directions: (1) broadening the source distribution for SFT data (more teachers, more styles) can mitigate the steepest collapse, and (2) RL without KL penalties can partially reverse DPO-induced semantic narrowing, though the effect is modest. Future work should investigate reasoning-path diversity (as distinct from final-answer diversity), test data-composition interventions directly, and examine whether the diversity floor we observe can be lowered by changes to the preference-optimization objective.

Acknowledgments

We would like to thank Samuel Lewis-Lim for his valuable feedback. CK is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation grant [grant number EP/S023062/1]. XT and NA are supported by the EPSRC [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project. We acknowledge (1) IT Services at the University of Sheffield for the provision of services for high-performance computing; (2) the use of the University of Oxford Advanced Research Computing (ARC) facility; (3) the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium through an EuroHPC Development Access call; (4) the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Creativity and Cognition*, pp. 413–425. ACM, June 2024. doi: 10.1145/3635636.3656204. URL http://dx.doi.org/10.1145/3635636.3656204. 2

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732. 3

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. Maxmin-RLHF: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=8tzjEMF0Vq. 1, 3

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374. 1, 3

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168. 3

Xingyu Dang, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Assessing diversity collapse in reasoning. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL https://openreview.net/forum?id=AMiKsHLjQh. 1, 2

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/. 3

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1. 4, 19

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=9rwtezthwo. 3

Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16568–16621. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gu24c.html. 3

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178. 31

Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. UniXcoder: Unified cross-modal pre-training for code representation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7212–7225, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.499. URL https://aclanthology.org/2022.acl-long.499/. 4, 19

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui

Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025a. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z. 3

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025b. doi: 10.1162/tacl.a.47. URL https://aclanthology.org/2025.tacl-1.69/. 2, 4

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. KL-regularized reinforcement learning is designed to mode collapse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=flBRtdIihA. 1, 2

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL https://github.com/huggingface/lighteval. 19

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe. 3

Shomik Jain, Jack Lanchantin, Maximilian Nickel, Karen Ullrich, Ashia Wilson, and Jamelle Watson-Daniels. Llm output homogenization is task dependent, 2025. URL https://arxiv.org/abs/2509.21267. 2

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz. 1, 2

Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. Diversity explains inference scaling laws: Through a case study of minimum Bayes risk decoding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29060–29094, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1410. URL https://aclanthology.org/2025.acl-long.1410/. 1

Constantinos Karouzos, Xingwei Tan, and Nikolaos Aletras. An empirical study on preference tuning generalization and diversity under domain shift. *arXiv preprint arXiv:2601.05882*, 2026. 2

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=DFr5hteojx. 3

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=PXD3FAVHJT. 1, 2, 20

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180. 19

Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6794–6814, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.346. URL https://aclanthology.org/2025.naacl-long.346/. 7

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH. 4, 31

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization, 2025. URL https://arxiv.org/abs/2501.18101. 3

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29838–29853, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1516. URL https://aclanthology.org/2025.emnlp-main.1516/. 7

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations, 2025a. URL https://arxiv.org/abs/2509.02534. 3

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=KfTf9vFvSn. 20, 27

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=NQEe7B7bSw. 3

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MKEHCx25xp. 3, 20, 28

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/. 3

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. Rethinking and refining the distinct metric. In Smaranda Muresan, Preslav Nakov, and

Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.86. URL https://aclanthology.org/2022.acl-short.86/. 4, 19

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692. 4, 19

Li-Chun Lu, Miri Liu, Pin Chun Lu, Yufei Tian, Shao-Hua Sun, and Nanyun Peng. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6329–6352, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.297. URL https://aclanthology.org/2026.eacl-long.297/. 20

Qinwei Ma, Jingzhe Shi, Can Jin, Jenq-Neng Hwang, Serge Belongie, and Lei Li. Gradient imbalance in direct preference optimization, 2025a. URL https://arxiv.org/abs/2502.20847. 1, 2

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking, 2025b. URL https://arxiv.org/abs/2504.09858. 1

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028/. 3

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/. 3

NVIDIA. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for Agentic reasoning, 2025. URL https://research.nvidia.com/labs/nemotron/files/NVIDIA-Nemotron-3-Nano-Technical-Report.pdf. Technical report. 31

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961. 2, 31

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=3pDMYjpOxk. 1

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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. 3

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155. 1

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL

https://openreview.net/forum?id=Feiz5HtCD0. 1, 2

Haihui Pan, Yuzhong Hong, Shaoke Lv, Junwei Bao, Hongfei Jiang, and Yang Song. Quality-constrained entropy maximization policy optimization for llm diversity, 2026. URL https://arxiv.org/abs/2602.15894. 3, 9

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Mind the gap: Conformative decoding to improve output diversity of instruction-tuned large language models, 2025. URL https://arxiv.org/abs/2507.20956. 1, 2, 3

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9. 1

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uaMSBJDnRv. 2

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/. 4, 19

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024. 3

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of LLM generated content. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O7bF6nlSOD. 2, 4, 20

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pOq9vDIYev. 3

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n. 1, 8

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=w6nlcS8Kkn. 7

Katherine Stasaski and Marti Hearst. Semantic diversity in dialogue with natural language inference. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 85–98, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.6. URL https://aclanthology.org/2022.naacl-main.6/. 4, 19

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 3

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, 2024. 3

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25/. 2

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024. 31

Alexandre Verine, Florian Le Bronnec, Kunhao Zheng, Alexandre Allauzen, Yann Chevaleyre, and benjamin negrevergne. Improving diversity in language models: When temperature fails, change the loss. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RsyMfsqzeG. 2

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://aclanthology.org/W17-4508/. 3

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2cRzmWXK9N. 1, 3, 9

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. 1, 7

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J. 2, 3

Peter West and Christopher Potts. Base models beat aligned models at randomness and creativity. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vqN8uom4A1. 2

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. Epistemic diversity and knowledge collapse in large language models, 2026. URL https://arxiv.org/abs/2510.04226. 1

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024. 2

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5. 2

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format: Diversity collapse in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15454–15468, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.836. URL https://aclanthology.org/2025.findings-emnlp.836/. 3

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM. 31

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao. 20, 27

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911. 3

A Implementation details

We generate outputs using vLLM (Kwon et al., 2023) and lighteval (Habib et al., 2023). For each model–task pair, we sample K=16 outputs per prompt (N=500 prompts; full dataset for Math-Geometry, IFEval, HumanEval, and TruthfulQA) with a 32,768-token generation length. All four diversity metrics (EAD, SBERT, NLI, Vendi Score) operate on the same post-stripping text. Table A lists all evaluation tasks with their sample sizes.

Category	Task	N	Category	Task	N	Category	Task	N
Summarization	TL;DR	500	Reasoning	GSM8K	500	Code	HumanEval	164
	CNN/DM	500		MATH-Alg	500		MBPP	500
	XSum	500		MATH-Geo	479		CRUXEval	500
Instruction	Alpaca	500		TruthfulQA	817	Value plur.	PRISM	500
	IFEval	541	Creative	WrtPrompts	500		WildBench	500

Table 2: Evaluation tasks grouped by category.

B Metric definitions

For a given prompt, the model generates K outputs $\{o_1, \ldots, o_K\}$. All metrics are computed per prompt and then averaged over prompts.

EAD (lexical diversity)

Expectation-Adjusted Distinct n-grams (Liu et al., 2022) counts the number of unique n-grams in the output set, normalized by the expected number of unique n-grams under a uniform draw from a vocabulary of size V. For a total of T n-gram tokens with U unique types, $\mathrm{EAD}_n = \frac{U}{V \cdot \left(1 - \left(\frac{V-1}{V}\right)^T\right)}$, where V is auto-detected from the model's tokenizer vocabulary. The denominator corrects for length bias: longer outputs are expected to contain more unique n-grams by chance. We average across $n \in \{1, \ldots, 5\}$ and clip to $[0, 1]$: $D_{\mathrm{EAD}} = \frac{1}{5} \sum_{n=1}^{5} \mathrm{EAD}_n$.

SBERT (semantic diversity)

We encode each output o_i with `all-mpnet-base-v2` (Reimers & Gurevych, 2019) to obtain L2-normalized embeddings $\mathbf{e}_i$. Semantic diversity is the mean pairwise cosine distance: $D_{\mathrm{SBERT}} = 1 - \frac{2}{K(K-1)} \sum_{i<j} \cos(\mathbf{e}_i, \mathbf{e}_j)$. Values near 0 indicate semantic collapse (all outputs map to the same region of embedding space); values near 1 indicate highly dissimilar outputs. For code tasks we additionally report diversity using UniXcoder (Guo et al., 2022), a code-aware encoder that captures structural similarity beyond surface tokens.

NLI (logical diversity)

Following Stasaski & Hearst (2022), we score output pairs with a natural language inference classifier (`roberta-large-mnli`; Liu et al., 2019). For each ordered pair (o_i, o_j), the model predicts a probability distribution over {entailment, neutral, contradiction}. We compute a directional similarity score as $P(\text{entailment}) - P(\text{contradiction})$, then symmetrize by averaging both orderings: $s_{ij} = \frac{1}{2}\left[\left(P_{\mathrm{ent}}(o_i \mid o_j) - P_{\mathrm{con}}(o_i \mid o_j)\right) + \left(P_{\mathrm{ent}}(o_j \mid o_i) - P_{\mathrm{con}}(o_j \mid o_i)\right)\right]$. Since NLI models are trained on single sentences rather than full paragraphs, we align sentences by position across outputs. The diversity score is: $D_{\mathrm{NLI}} = 1 - \frac{2}{K(K-1)} \sum_{i<j} s_{ij}$. D_{NLI} near 0 indicates mutual entailment (collapse), near 1 indicates neutrality, and values above 1 indicate net contradiction (the outputs make mutually inconsistent claims). Code tasks are excluded as NLI is not meaningful for program text.

Vendi Score

The Vendi Score (Friedman & Dieng, 2023) measures the effective number of dissimilar elements via the eigenvalue entropy of a similarity kernel. We reuse the SBERT cosine

similarity matrix. Given K outputs with L2-normalized embeddings, we form the Gram matrix $\mathbf{G}$ where $G_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$ and trace-normalize it as $\mathbf{P} = \mathbf{G}/K$. The Vendi Score is $\text{VS} = \exp(-\sum_i \lambda_i \log \lambda_i)$, where λ_i are the eigenvalues of $\mathbf{P}$. VS=1 when all outputs are identical (rank-1 kernel) and VS=K when all outputs are orthogonal (full-rank uniform spectrum). Because the Vendi Score shares the SBERT kernel, agreement between VS and D_{SBERT} is expected rather than independent confirmation; VS adds the interpretable "effective number of modes" framing.

AST subtree diversity (structural, code only)

For code-generation tasks (HumanEval, MBPP), we measure structural diversity via the mean pairwise Jaccard distance on AST subtree multisets (subtree height ≤ 4; Shypula et al., 2025). We parse each output into a Python AST, extract all subtrees up to height 4, represent each output as a multiset of subtree hashes, and compute $D_{\text{AST}}(o_i, o_j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$, where S_i is the multiset of subtree hashes for output o_i. This metric is reported on correct (executable, test-passing) outputs only, to capture genuine structural variation among working solutions. Unparseable outputs are excluded.

LLM-as-a-Judge quality

For the eight tasks without verifiable answers, we evaluate quality using established LLM-as-judge frameworks with `gpt-4.1-mini` via the OpenAI Batch API. *Summarization* (TL;DR, CNN/DM, XSum): pairwise win-rate against reference summaries, following Kirk et al. (2024b). *Instruction following and value pluralism* (Alpaca, PRISM): pairwise comparison against Base using MT-Bench prompts (Zheng et al., 2023). *Creative writing* (WritingPrompts): pairwise comparison using Arena-Hard creative writing prompts (Li et al., 2025b). *WildBench*: checklist-guided WB-Score (Lin et al., 2025). We note that LLM-judge evaluation of creative and value-laden tasks has known limitations (Lu et al., 2026); we report these results as supplementary context for our diversity findings rather than as primary evidence.

C Per-task diversity results

Tables 3–6 report per input diversity for each of the four metrics across all 15 tasks and 16 models (13 standard + 3 Think w/o CoT). Table 3 reports SBERT cosine distance, our primary semantic diversity measure. Table 4 reports Expected Agreement Diversity (EAD), a lexical overlap metric. Table 5 reports NLI-based diversity, which captures inferential disagreement between output pairs; code tasks are excluded as NLI is not meaningful for program text. Table 6 reports Vendi Score, the effective number of distinct semantic modes among the K=16 outputs.

D Quality results

Tables 7–13 report task performance for all 16 models, across all 15 tasks. Table 7 reports reasoning quality on four tasks (GSM8K, MATH-Algebra, MATH-Geometry, TruthfulQA) with accuracy@1, majority vote@16, and pass@16. Table 8 reports code generation quality (pass@k for $k \in \{1, 5, 10, 16\}$) on HumanEval and MBPP. Table 9 reports IFEval constraint satisfaction with strict and loose accuracy@1, pass@16, and consistency@16. Table 10 reports CruxEval output-prediction accuracy.

For tasks without verifiable answers, we use LLM-as-judge evaluation with gpt-4.1-mini via the OpenAI Batch API. Table 11 reports pairwise win rates against reference summaries following Kirk et al. (2024b). Table 12 reports pairwise win rates against the Base model using the MT-Bench prompt (Zheng et al., 2023) for Alpaca and PRISM and the Arena-Hard creative writing prompt (Li et al., 2025b) for WritingPrompts. Table 13 reports checklist-guided WB-Score (Lin et al., 2025). We note that LLM-judge evaluation of creative and value-laden tasks has known limitations (Lu et al., 2026); we report these results as supplementary context for our diversity findings rather than as primary evidence.

	Summarization			*Instruction F.*		*Creative Wr.*	*Value Pluralism*	
	TL;DR	CNN/DM	XSum	Alpaca	IFEval	WritingPrompts	PRISM	WildBench
Base	0.353	0.279	0.451	0.319	0.349	0.540	0.408	0.335
Instruct-SFT	0.268	0.223	0.282	0.170	0.172	0.276	0.141	0.129
Instruct-DPO	0.202	0.075	0.083	0.120	0.154	0.225	0.096	0.122
Instruct (final)	0.207	0.072	0.081	0.113	0.154	0.202	0.090	0.118
Think-SFT	0.168	0.083	0.090	0.141	0.191	0.240	0.100	0.160
Think-DPO	0.159	0.059	0.064	0.118	0.165	0.205	0.089	0.154
Think (final)	0.161	0.091	0.092	0.146	0.196	0.199	0.091	0.173
Think-SFT w/o CoT	0.293	0.249	0.202	0.137	0.196	0.266	0.114	0.191
Think-DPO w/o CoT	0.344	0.176	0.130	0.104	0.157	0.223	0.100	0.156
Think w/o CoT	0.323	0.220	0.167	0.161	0.221	0.245	0.102	0.181
RL-Zero-Math	0.336	0.201	0.436	0.309	0.318	0.543	0.393	0.313
RL-Zero-Code	0.327	0.193	0.422	0.178	0.287	0.533	0.367	0.262
RL-Zero-IF	0.333	0.210	0.429	0.176	0.397	0.546	0.400	0.300
RL-Zero-General	0.309	0.184	0.404	0.155	0.284	0.523	0.372	0.279
RL-Zero-Math$^{3.1}$	0.330	0.200	0.432	0.319	0.324	0.546	0.398	0.316
RL-Zero-Code$^{3.1}$	0.328	0.196	0.430	0.314	0.325	0.539	0.394	0.315

	Reasoning				*Code*		
	GSM8K	MATH-Alg	MATH-Geo	TruthfulQA	HumanEval	MBPP	CRUXEval
Base	0.172	0.146	0.198	0.353	0.411	0.291	0.239
Instruct-SFT	0.105	0.132	0.179	0.327	0.112	0.111	0.218
Instruct-DPO	0.141	0.071	0.096	0.158	0.095	0.073	0.068
Instruct (final)	0.078	0.057	0.101	0.115	0.093	0.069	0.062
Think-SFT	0.061	0.054	0.107	0.119	0.109	0.081	0.095
Think-DPO	0.052	0.061	0.114	0.074	0.081	0.084	0.076
Think (final)	0.051	0.062	0.122	0.075	0.117	0.089	0.090
Think-SFT w/o CoT	0.057	0.066	0.098	0.106	0.055	0.084	0.084
Think-DPO w/o CoT	0.045	0.058	0.077	0.085	0.062	0.083	0.064
Think w/o CoT	0.052	0.064	0.089	0.089	0.060	0.083	0.071
RL-Zero-Math	0.154	0.144	0.181	0.352	0.421	0.274	0.222
RL-Zero-Code	0.156	0.144	0.183	0.348	0.464	0.238	0.149
RL-Zero-IF	0.177	0.143	0.199	0.357	0.336	0.297	0.491
RL-Zero-General	0.133	0.124	0.166	0.326	0.468	0.272	0.198
RL-Zero-Math$^{3.1}$	0.183	0.140	0.183	0.358	0.460	0.292	0.207
RL-Zero-Code$^{3.1}$	0.173	0.139	0.178	0.349	0.439	0.261	0.209

Table 3: Per-input **SBERT** diversity (`all-mpnet-base-v2`).

	Summarization			*Instruction F.*		*Creative Wr.*	*Value Pluralism*	
	TL;DR	CNN/DM	XSum	Alpaca	IFEval	WritingPrompts	PRISM	WildBench
Base	0.37	0.37	0.67	0.51	0.44	0.23	0.24	0.30
Instruct-SFT	0.69	0.43	0.58	0.57	0.62	0.72	0.68	0.68
Instruct-DPO	0.71	0.53	0.51	0.56	0.71	0.80	0.72	0.76
Instruct (final)	0.68	0.50	0.48	0.52	0.67	0.79	0.70	0.74
Think-SFT	0.76	0.59	0.58	0.61	0.58	0.73	0.72	0.63
Think-DPO	0.79	0.62	0.63	0.69	0.74	0.83	0.76	0.75
Think (final)	0.78	0.59	0.59	0.65	0.68	0.80	0.74	0.71
Think-SFT w/o CoT	0.44	0.42	0.45	0.65	0.61	0.75	0.71	0.59
Think-DPO w/o CoT	0.70	0.56	0.61	0.67	0.72	0.81	0.73	0.69
Think w/o CoT	0.56	0.46	0.49	0.63	0.64	0.77	0.70	0.64
RL-Zero-Math	0.41	0.38	0.64	0.49	0.55	0.33	0.36	0.44
RL-Zero-Code	0.47	0.39	0.66	0.58	0.60	0.40	0.43	0.53
RL-Zero-IF	0.46	0.41	0.66	0.47	0.55	0.33	0.39	0.50
RL-Zero-General	0.45	0.38	0.65	0.62	0.62	0.36	0.44	0.51
RL-Zero-Math$^{3.1}$	0.34	0.36	0.62	0.53	0.53	0.30	0.35	0.43
RL-Zero-Code$^{3.1}$	0.35	0.36	0.61	0.56	0.54	0.29	0.35	0.45

	Reasoning				*Code*		
	GSM8K	MATH-Alg	MATH-Geo	TruthfulQA	HumanEval	MBPP	CRUXEval
Base	0.45	0.45	0.40	0.46	0.57	0.59	0.31
Instruct-SFT	0.38	0.45	0.51	0.57	0.48	0.51	0.57
Instruct-DPO	0.47	0.44	0.56	0.65	0.57	0.57	0.58
Instruct (final)	0.36	0.39	0.52	0.64	0.55	0.54	0.55
Think-SFT	0.36	0.30	0.41	0.62	0.43	0.43	0.48
Think-DPO	0.36	0.32	0.46	0.64	0.42	0.48	0.50
Think (final)	0.32	0.32	0.45	0.61	0.46	0.46	0.50
Think-SFT w/o CoT	0.41	0.43	0.50	0.69	0.40	0.54	0.52
Think-DPO w/o CoT	0.37	0.41	0.49	0.72	0.46	0.58	0.51
Think w/o CoT	0.40	0.43	0.49	0.67	0.45	0.55	0.50
RL-Zero-Math	0.41	0.49	0.49	0.57	0.59	0.61	0.41
RL-Zero-Code	0.47	0.49	0.51	0.58	0.56	0.60	0.41
RL-Zero-IF	0.46	0.43	0.50	0.57	0.61	0.64	0.55
RL-Zero-General	0.45	0.45	0.47	0.52	0.57	0.59	0.45
RL-Zero-Math$^{3.1}$	0.41	0.46	0.48	0.54	0.57	0.60	0.40
RL-Zero-Code$^{3.1}$	0.43	0.47	0.49	0.54	0.56	0.59	0.43

Table 4: Per-input **EAD** diversity.

	Summarization			*Instruction. F.*		*Creative Wr.*	*Value Pluralism*	
	TL;DR	CNN/DM	XSum	Alpaca	IFEval	WritingPrompts	PRISM	WildBench
Base	0.95	1.04	1.09	0.68	1.05	1.16	1.09	1.06
Instruct-SFT	0.90	0.71	0.99	0.78	0.97	1.02	0.93	1.05
Instruct-DPO	0.86	0.84	0.77	0.77	0.98	1.05	0.97	1.06
Instruct (final)	0.84	0.79	0.72	0.73	0.93	1.02	0.95	1.05
Think-SFT	1.02	0.93	0.92	0.89	1.01	1.13	1.04	1.09
Think-DPO	1.06	0.93	0.93	0.93	1.06	1.18	1.07	1.12
Think (final)	1.03	0.90	0.89	0.85	1.00	1.12	1.04	1.09
Think-SFT w/o CoT	1.04	0.98	0.99	0.96	1.01	1.12	1.04	1.10
Think-DPO w/o CoT	0.98	0.96	0.97	0.99	1.06	1.18	1.08	1.10
Think w/o CoT	1.00	0.98	0.97	0.91	1.00	1.09	1.02	1.09
RL-Zero-Math	0.92	0.90	1.05	0.69	1.05	1.16	1.09	1.08
RL-Zero-Code	0.90	0.89	1.04	0.97	1.05	1.14	1.07	1.08
RL-Zero-IF	0.89	0.85	1.04	0.68	0.89	1.15	1.06	1.01
RL-Zero-General	0.89	0.89	1.04	0.85	1.02	1.14	1.06	1.06
RL-Zero-Math$^{3.1}$	0.92	0.90	1.05	0.69	1.05	1.15	1.08	1.07
RL-Zero-Code$^{3.1}$	0.91	0.89	1.05	0.74	1.06	1.15	1.08	1.07

	Reasoning			
	GSM8K	MATH-Alg	MATH-Geo	TruthfulQA
Base	1.08	1.00	1.13	0.97
Instruct-SFT	0.77	1.01	1.10	0.88
Instruct-DPO	0.77	0.76	0.88	0.91
Instruct (final)	0.73	0.76	0.89	0.90
Think-SFT	0.77	0.72	0.85	0.98
Think-DPO	0.73	0.72	0.86	0.98
Think (final)	0.70	0.73	0.86	0.99
Think-SFT w/o CoT	0.90	0.87	1.00	1.03
Think-DPO w/o CoT	0.81	0.90	1.02	1.05
Think w/o CoT	0.87	0.91	1.03	1.00
RL-Zero-Math	1.05	0.99	1.09	0.97
RL-Zero-Code	1.05	0.98	1.09	0.96
RL-Zero-IF	1.01	0.96	1.10	0.95
RL-Zero-General	1.02	0.95	1.08	0.94
RL-Zero-Math$^{3.1}$	1.06	0.98	1.10	0.97
RL-Zero-Code$^{3.1}$	1.05	0.98	1.09	0.97

Table 5: Per-input **NLI** diversity. Code tasks excluded.

	Summarization			*Instruction F.*		*Creative Wr.*	*Value Pluralism*	
	TL;DR	CNN/DM	XSum	Alpaca	IFEval	WritingPrompts	PRISM	WildBench
Base	4.2	3.2	5.2	2.2	3.8	6.9	4.6	3.5
Instruct-SFT	3.0	2.4	3.2	2.1	2.3	3.2	2.0	1.9
Instruct-DPO	2.4	1.5	1.6	1.8	2.1	2.8	1.7	1.9
Instruct (final)	2.5	1.5	1.5	1.7	2.2	2.6	1.6	1.8
Think-SFT	2.2	1.6	1.6	2.0	2.4	2.9	1.7	2.2
Think-DPO	2.2	1.4	1.4	1.9	2.3	2.6	1.6	2.1
Think (final)	2.2	1.6	1.6	2.0	2.5	2.6	1.6	2.3
Think-SFT w/o CoT	3.0	2.5	2.2	2.0	2.4	3.1	1.8	2.3
Think-DPO w/o CoT	2.8	2.0	1.8	1.7	2.1	2.7	1.6	2.0
Think w/o CoT	3.0	2.3	2.0	2.0	2.4	2.8	1.7	2.2
RL-Zero-Math	3.9	2.4	4.9	2.3	3.5	7.0	4.5	3.3
RL-Zero-Code	3.8	2.3	4.7	2.1	3.2	6.8	4.2	2.9
RL-Zero-IF	3.8	2.4	4.8	2.0	4.4	7.0	4.5	3.1
RL-Zero-General	3.6	2.3	4.5	2.0	3.2	6.7	4.2	3.0
RL-Zero-Math$^{3.1}$	3.8	2.4	4.8	2.2	3.6	7.0	4.6	3.3
RL-Zero-Code$^{3.1}$	3.8	2.4	4.8	2.3	3.5	6.9	4.5	3.3

	Reasoning				*Code*		
	GSM8K	MATH-Alg	MATH-Geo	TruthfulQA	HumanEval	MBPP	CRUXEval
Base	2.1	2.0	2.4	3.8	2.5	2.8	2.5
Instruct-SFT	1.7	1.9	2.3	3.4	1.7	1.7	2.2
Instruct-DPO	1.9	1.5	1.6	2.0	1.7	1.5	1.5
Instruct (final)	1.5	1.4	1.7	1.7	1.6	1.5	1.4
Think-SFT	1.4	1.4	1.7	1.8	1.6	1.5	1.6
Think-DPO	1.3	1.4	1.8	1.5	1.5	1.6	1.5
Think (final)	1.3	1.4	1.8	1.5	1.7	1.6	1.6
Think-SFT w/o CoT	1.4	1.4	1.6	1.7	1.4	1.6	1.6
Think-DPO w/o CoT	1.3	1.4	1.5	1.6	1.4	1.5	1.4
Think w/o CoT	1.3	1.4	1.6	1.6	1.4	1.5	1.5
RL-Zero-Math	2.0	2.0	2.3	3.8	2.6	2.7	2.3
RL-Zero-Code	2.0	2.0	2.3	3.7	3.0	2.5	1.9
RL-Zero-IF	2.1	1.9	2.4	3.8	2.0	2.7	3.9
RL-Zero-General	1.9	1.8	2.2	3.5	2.9	2.6	2.2
RL-Zero-Math$^{3.1}$	2.2	1.9	2.3	3.9	2.9	2.8	2.1
RL-Zero-Code$^{3.1}$	2.1	1.9	2.2	3.8	2.7	2.6	2.1

Table 6: Per-input **Vendi Score** diversity.

	GSM8K			MATH-Algebra			MATH-Geometry			TruthfulQA		
	acc	mv	pass	acc	mv	pass	acc	mv	pass	acc	mv	pass
Base	56.0	80.4	94.8	50.0	59.4	75.6	20.5	24.6	50.3	10.0	8.6	28.5
Instruct-SFT	73.4	84.4	95.4	56.2	68.2	**83.2**	26.5	36.7	61.0	9.4	7.2	24.0
Instruct-DPO	77.2	86.4	96.2	51.4	65.8	81.6	23.0	35.7	54.9	8.2	6.7	20.8
Instruct (final)	80.4	87.6	95.2	70.8	75.0	81.2	42.6	54.3	**63.3**	8.1	8.0	19.6
Think-SFT	92.0	**93.4**	**97.0**	76.4	77.2	78.8	50.5	54.7	59.5	9.7	7.0	21.5
Think-DPO	85.2	89.4	95.6	74.6	77.0	78.2	50.5	54.5	61.6	7.0	6.6	13.8
Think (final)	**93.0**	**93.4**	96.4	**76.8**	**77.6**	78.8	**51.1**	**55.3**	59.7	8.6	6.9	19.3
Think-SFT w/o CoT	76.6	82.4	94.6	56.4	63.8	75.0	27.6	29.9	46.8	8.6	7.6	16.3
Think-DPO w/o CoT	70.0	79.4	94.6	47.4	52.8	67.8	19.6	23.2	37.2	7.1	5.5	12.1
Think w/o CoT	74.6	82.8	94.0	48.6	55.4	68.2	19.6	25.5	38.8	9.4	7.8	16.8
RL-Zero-Math	61.0	83.2	95.8	49.4	64.6	79.2	22.8	27.3	57.6	9.7	7.6	29.4
RL-Zero-Code	58.2	83.8	96.4	51.2	63.4	80.8	23.0	28.2	57.8	**12.2**	8.2	**30.6**
RL-Zero-IF	49.8	75.0	94.2	48.2	60.8	77.4	21.3	25.5	51.6	10.2	**9.3**	28.8
RL-Zero-General	61.0	82.8	**97.0**	54.0	64.8	79.6	24.4	29.6	57.2	10.9	7.7	28.8
RL-Zero-Math$^{3.1}$	55.2	80.6	96.2	53.6	66.0	78.8	20.9	28.4	55.7	10.6	8.4	28.2
RL-Zero-Code$^{3.1}$	59.8	81.8	95.2	52.4	62.8	80.4	22.1	27.1	56.8	10.2	8.2	27.5

Table 7: **Reasoning** quality (%). acc: first correct. mv: majority vote. pass: any of K=16 correct.

	HumanEval				MBPP			
	@1	@5	@10	@16	@1	@5	@10	@16
Base	1.6	6.6	10.8	14.0	23.9	45.0	50.9	54.0
Instruct-SFT	63.4	88.1	93.6	96.3	32.3	47.5	52.1	54.8
Instruct-DPO	73.3	93.6	96.4	97.0	32.9	47.9	51.8	53.6
Instruct (final)	81.2	**96.2**	**97.7**	**98.2**	37.8	48.9	51.7	53.2
Think-SFT	86.7	94.9	95.6	95.7	41.0	50.1	52.3	53.6
Think-DPO	86.5	94.5	95.0	95.1	40.6	49.7	51.9	52.8
Think (final)	**87.7**	95.0	95.6	95.7	**44.1**	**53.7**	**56.1**	**58.0**
Think-SFT w/o CoT	49.4	76.3	81.8	84.1	24.0	43.2	48.6	51.4
Think-DPO w/o CoT	56.5	78.4	82.4	84.8	26.2	42.9	47.2	49.4
Think w/o CoT	55.6	77.6	82.0	84.1	23.9	42.1	47.7	50.8
RL-Zero-Math	2.4	10.3	17.1	23.2	24.5	45.6	51.7	55.0
RL-Zero-Code	2.7	11.2	19.1	26.8	24.8	44.8	50.2	53.2
RL-Zero-IF	1.1	5.1	9.4	13.4	24.6	45.0	51.8	56.0
RL-Zero-General	2.5	11.1	19.7	28.0	24.9	44.9	51.0	54.6
RL-Zero-Math$^{3.1}$	2.1	9.0	15.5	21.3	24.1	44.7	50.2	53.0
RL-Zero-Code$^{3.1}$	66.5	83.9	87.9	89.6	25.4	45.4	50.6	53.2

Table 8: **Code** quality (pass@k, %).

	strict@1	loose@1	pass@16	consist
Base	44.7	58.4	74.1	46.5
Instruct-SFT	78.7	85.9	90.6	79.3
Instruct-DPO	78.7	85.3	89.6	79.3
Instruct (final)	**82.1**	**87.9**	89.3	**81.8**
Think-SFT	78.0	84.9	90.9	77.2
Think-DPO	74.9	81.4	86.7	73.7
Think (final)	78.7	85.2	**91.7**	79.5
Think-SFT w/o CoT	70.2	79.3	89.5	71.5
Think-DPO w/o CoT	66.7	75.4	82.4	66.5
Think w/o CoT	71.0	79.6	88.0	70.7
RL-Zero-Math	47.7	59.8	72.6	47.0
RL-Zero-Code	47.0	60.1	71.5	46.6
RL-Zero-IF	59.7	70.5	75.0	61.0
RL-Zero-General	46.6	59.6	72.3	48.0
RL-Zero-Math$^{3.1}$	48.6	60.7	72.5	45.9
RL-Zero-Code$^{3.1}$	46.4	58.5	73.6	46.7

Table 9: **IFEval** constraint satisfaction (%).

	Acc@1	MV@16	Pass@16
Base	16.4	36.0	61.0
Instruct-SFT	32.4	43.5	73.6
Instruct-DPO	**32.9**	40.0	**84.5**
Instruct (final)	18.0	21.0	76.2
Think-SFT	19.2	28.2	74.5
Think-DPO	17.4	26.5	65.1
Think (final)	15.8	29.4	65.5
Think-SFT w/o CoT	26.3	**47.5**	74.2
Think-DPO w/o CoT	27.0	44.2	70.7
Think w/o CoT	27.7	44.2	71.4
RL-Zero-Math	14.2	34.9	59.8
RL-Zero-Code	10.0	27.0	52.2
RL-Zero-IF	23.0	32.9	58.8
RL-Zero-General	18.8	37.9	68.2
RL-Zero-Math$^{3.1}$	9.8	25.8	50.5
RL-Zero-Code$^{3.1}$	10.8	28.4	55.0

Table 10: CruxEval output prediction quality (%). Accuracy@1, majority vote@16, and pass@16.

	TL;DR	CNN/DM	XSum
Base	26.0	47.7	26.7
Instruct-SFT	72.2	20.0	52.0
Instruct-DPO	70.4	95.4	94.4
Instruct (final)	**77.8**	95.4	95.4
Think-SFT	32.0	97.2	95.8
Think-DPO	28.4	91.6	90.6
Think (final)	38.2	**97.8**	**96.0**
Think-SFT w/o CoT	20.0	55.6	78.4
Think-DPO w/o CoT	13.2	44.7	67.8
Think w/o CoT	12.5	49.6	73.9
RL-Zero-Math	35.4	49.3	37.4
RL-Zero-Code	37.2	49.0	41.4
RL-Zero-IF	36.8	41.6	39.4
RL-Zero-General	44.0	60.6	43.9
RL-Zero-Math$^{3.1}$	34.4	50.0	39.8
RL-Zero-Code$^{3.1}$	33.0	56.2	40.6

Table 11: Summarization quality: pairwise win rate (%) against reference summaries, judged by gpt-4.1-mini.

	Alpaca	PRISM	WritingPrompts
Base	—	—	—
Instruct-SFT	48.7	84.0	93.1
Instruct-DPO	73.9	93.1	96.9
Instruct (final)	66.3	91.0	97.3
Think-SFT	84.3	92.5	96.8
Think-DPO	**95.2**	**95.7**	**98.0**
Think (final)	83.9	93.7	97.3
Think-SFT w/o CoT	83.4	88.6	92.5
Think-DPO w/o CoT	88.0	92.0	95.4
Think w/o CoT	84.7	88.6	93.3
RL-Zero-Math	53.1	51.7	49.9
RL-Zero-Code	55.4	61.9	54.6
RL-Zero-IF	32.8	53.3	52.4
RL-Zero-General	76.7	63.0	53.6
RL-Zero-Math$^{3.1}$	57.5	55.7	49.2
RL-Zero-Code$^{3.1}$	52.5	52.1	49.4

Table 12: Open-ended quality: pairwise win rate (%) against Base model. Alpaca and PRISM use the MT-Bench pair-v2 prompt (Zheng et al., 2023); WritingPrompts uses the Arena-Hard creative writing prompt (Li et al., 2025b) with position-swap debiasing. Judge: gpt-4.1-mini.

	Raw	σ	Median	WB-Score
Base	4.0	2.4	4	-2.0
Instruct-SFT	7.2	1.9	8	4.5
Instruct-DPO	7.6	1.7	8	5.2
Instruct (final)	**8.0**	1.5	**9**	**6.1**
Think-SFT	7.2	2.1	8	4.3
Think-DPO	7.5	2.0	8	5.1
Think (final)	7.3	2.0	8	4.6
Think-SFT w/o CoT	5.4	2.6	5	0.8
Think-DPO w/o CoT	5.7	2.3	6	1.4
Think w/o CoT	5.7	2.5	6	1.4
RL-Zero-Math	4.1	2.5	4	-1.7
RL-Zero-Code	4.2	2.6	4	-1.6
RL-Zero-IF	4.0	2.5	4	-2.0
RL-Zero-General	4.9	2.7	5	-0.2
RL-Zero-Math$^{3.1}$	4.0	2.5	4	-2.0
RL-Zero-Code$^{3.1}$	4.2	2.6	4	-1.6

Table 13: WildBench quality: checklist-guided WB-Score (Lin et al., 2025), judged by gpt-4.1-mini. Raw score (1–10) and normalized WB-Score $= (\text{raw} - 5) \times 2$.

E Quality-filtered diversity

Table 14 reports the quality-filtered diversity decomposition defined in §3.3 for six verifiable tasks. We label each of K=16 generations as correct or incorrect (answer matching for math, test execution for code, constraint satisfaction for IFEval), then report accuracy alongside D_a (SBERT on all outputs), D_c (SBERT on correct-only subset, $K_c \geq 2$), and V_c (Vendi Score on correct outputs, interpreted as the effective number of distinct correct answers).

F Code-specific diversity

Table 15 reports quality-filtered code diversity using the domain-specific metrics described in §3.3: UniXcoder SBERT (D_c^{code}, computed on correct outputs only) and AST subtree Jaccard distance (D_c^{AST}, for code-generation tasks). Missing entries ("—") indicate models with no parseable correct outputs.

	GSM8K				MATH-Algebra				MATH-Geometry			
	acc	D_a	D_c	V_c	acc	D_a	D_c	V_c	acc	D_a	D_c	V_c
Base	52	0.172	0.135	1.7	48	0.146	0.119	1.6	23	0.198	0.145	1.6
Instruct-SFT	73	0.105	0.098	1.5	56	0.132	0.110	1.6	26	0.179	0.140	1.6
Instruct-DPO	77	0.141	0.137	1.8	51	0.071	0.067	1.4	23	0.096	0.082	1.4
Instruct (final)	80	0.078	0.074	1.4	71	0.057	0.057	1.4	43	0.101	0.087	1.5
Think-SFT	92	0.061	0.060	1.4	76	0.054	0.051	1.3	50	0.107	0.080	1.5
Think-DPO	85	0.052	0.049	1.3	75	0.061	0.053	1.3	50	0.114	0.082	1.5
Think (final)	**93**	0.051	0.050	1.3	**77**	0.062	0.059	1.4	**51**	0.122	0.091	1.6
Think-SFT w/o CoT	77	0.057	0.055	1.3	56	0.066	0.058	1.3	28	0.098	0.072	1.4
Think-DPO w/o CoT	70	0.045	0.042	1.2	47	0.058	0.050	1.3	20	0.077	0.061	1.3
Think w/o CoT	75	0.052	0.048	1.3	49	0.064	0.055	1.3	20	0.089	0.064	1.3
RL-Zero-Math	61	0.154	0.124	1.7	49	0.144	0.119	1.6	23	0.181	0.135	1.6
RL-Zero-Code	58	0.156	0.127	1.7	51	0.144	0.114	1.6	23	0.183	0.135	1.6
RL-Zero-IF	50	0.177	0.137	1.7	48	0.143	0.111	1.6	21	0.199	0.132	1.6
RL-Zero-General	61	0.133	0.110	1.6	54	0.124	0.104	1.6	24	0.166	0.127	1.5
RL-Zero-Math$^{3.1}$	55	0.183	0.136	1.7	54	0.140	0.120	1.6	21	0.183	0.133	1.6
RL-Zero-Code$^{3.1}$	60	0.173	0.130	1.7	52	0.139	0.115	1.6	22	0.178	0.133	1.6

	IFEval				HumanEval				MBPP				CRUXEval			
	acc	D_a	D_c	V_c	acc	D_a	D_c	V_c	acc	D_a	D_c	V_c	acc	D_a	D_c	V_c
Base	45	0.349	0.333	3.2	18	0.411	0.123	1.5	19	0.291	0.196	1.9	20	0.239	0.240	1.9
Instruct-SFT	79	0.172	0.171	2.2	63	0.112	0.109	1.6	32	0.111	0.098	1.5	32	0.218	0.177	1.7
Instruct-DPO	79	0.154	0.155	2.1	73	0.095	0.095	1.6	33	0.073	0.059	1.3	**38**	0.068	0.168	1.6
Instruct (final)	**82**	0.154	0.155	2.1	81	0.093	0.091	1.6	38	0.069	0.058	1.3	23	0.062	0.139	1.4
Think-SFT	78	0.191	0.180	2.3	87	0.109	0.101	1.6	41	0.081	0.058	1.3	18	0.095	0.076	1.3
Think-DPO	75	0.165	0.159	2.1	87	0.081	0.072	1.4	36	0.084	0.067	1.4	17	0.076	0.056	1.2
Think (final)	79	0.196	0.187	2.3	**88**	0.117	0.110	1.6	**44**	0.089	0.064	1.4	12	0.090	0.074	1.3
Think-SFT w/o CoT	70	0.196	0.185	2.2	49	0.055	0.046	1.3	24	0.084	0.072	1.4	20	0.084	0.087	1.4
Think-DPO w/o CoT	67	0.157	0.152	2.0	56	0.062	0.051	1.3	26	0.083	0.065	1.3	21	0.064	0.098	1.4
Think w/o CoT	71	0.221	0.182	2.1	56	0.060	0.053	1.3	24	0.083	0.070	1.4	20	0.071	0.081	1.3
RL-Zero-Math	48	0.318	0.295	2.9	3	0.421	0.089	1.4	24	0.274	0.157	1.8	18	0.222	0.245	1.9
RL-Zero-Code	47	0.287	0.278	2.7	3	0.464	0.180	1.5	25	0.238	0.147	1.7	16	0.149	0.201	1.7
RL-Zero-IF	60	0.397	0.371	3.9	0	0.336	–	–	24	0.297	0.149	1.7	24	0.491	0.319	2.1
RL-Zero-General	47	0.284	0.271	2.7	32	0.468	0.113	1.5	25	0.272	0.151	1.7	23	0.198	0.190	1.8
RL-Zero-Math$^{3.1}$	49	0.324	0.300	2.9	7	0.460	0.116	1.4	24	0.292	0.157	1.8	19	0.207	0.236	1.8
RL-Zero-Code$^{3.1}$	46	0.325	0.293	2.8	66	0.439	0.071	1.4	26	0.261	0.153	1.8	17	0.209	0.247	1.8

Table 14: Quality-filtered diversity. acc: accuracy (%). D_a: SBERT on all outputs. D_c: SBERT on correct only ($K_c \geq 2$). V_c: Vendi Score on correct only (effective number of distinct answers).

	HumanEval			MBPP		
	acc	D_c^{code}	D_c^{AST}	acc	D_c^{code}	D_c^{AST}
Base	18	0.168	0.590	19	0.310	0.927
Instruct-SFT	63	0.167	0.591	32	0.179	0.683
Instruct-DPO	74	0.113	0.674	33	0.118	0.777
Instruct (final)	81	0.142	0.593	38	0.118	0.693
Think-SFT	**91**	0.116	0.527	40	0.124	0.533
Think-DPO	**91**	0.130	0.540	39	0.162	0.611
Think (final)	**91**	0.126	0.531	**44**	0.130	0.590
Think-SFT w/o CoT	51	0.105	0.510	26	0.170	0.662
Think-DPO w/o CoT	59	0.123	0.485	28	0.160	0.707
Think w/o CoT	57	0.112	0.499	26	0.164	0.676
RL-Zero-Math	3	0.058	0.057	25	0.253	0.905
RL-Zero-Code	3	0.254	—	25	0.249	0.889
RL-Zero-IF	0	—	—	25	0.245	0.887
RL-Zero-General	33	0.174	0.618	25	0.255	0.900
RL-Zero-Math$^{3.1}$	7	0.101	0.261	24	0.263	0.889
RL-Zero-Code$^{3.1}$	67	0.124	0.576	25	0.249	0.895

Table 15: Code-specific diversity on correct outputs for code-generation tasks. acc: accuracy (%, mean $K_c/16$). D_c^{code}: UniXcoder SBERT (correct only). D_c^{AST}: AST subtree Jaccard (correct only).

G Output length analysis

Table G reports the mean output word length and mean SBERT diversity per task, averaged across all 13 models. Tasks with high mean diversity (e.g. WritingPrompts, HumanEval) span a wide range of output lengths, and tasks with similar lengths (e.g. GSM8K at 137 words, TruthfulQA at 142 words) have very different diversity levels (0.128 vs. 0.262). Output length does not systematically predict diversity.

Task	Len	SBERT	Task	Len	SBERT
WildBench	872	0.230	TL;DR	283	0.270
PRISM	723	0.260	MATH-Algebra	227	0.110
WritingPrompts	704	0.397	MBPP	213	0.198
CRUXEval	619	0.183	HumanEval	211	0.280
MATH-Geometry	441	0.155	XSum	158	0.292
IFEval	391	0.257	TruthfulQA	142	0.262
Alpaca	304	0.204	GSM8K	137	0.128
			CNN/DailyMail	120	0.157

Table 16: Mean output word length and SBERT diversity per task, averaged across 13 models.

H Temperature sensitivity

Table 17 compares Base model diversity at its recommended sampling temperature (T=1.0, top-p=0.7) with the matched temperature used throughout this study (T=0.6, top-p=0.95). SBERT diversity decreases by 11% on average, EAD by 18%, and NLI by only 3%. These reductions are modest relative to the 62% SBERT drop from Base to Think-SFT, confirming that the diversity gaps documented in this paper are not attributable to the temperature difference.

I Stage attribution per task

Table 18 reports the percentage of Base SBERT diversity lost at each post-training stage for all 15 tasks. Think collapses 45–80% at SFT (most on XSum, least on IFEval), with DPO contributing minimally. Instruct shows the opposite pattern: SFT losses range from 8–73%, but DPO contributes 2–63% additional loss. RL-Zero retains 71–105% of Base diversity across tasks.

J Decontamination

We measure training–evaluation data overlap using C_{13} 13-gram matching (Lambert et al., 2025): for each test instance, we extract all 13-grams (tokenized with spaCy), query an Elasticsearch index of the training data for phrase matches, and report the fraction of test tokens covered by at least one matching 13-gram, averaged over all test instances. Table 19 reports results for the four Dolci post-training datasets against all fifteen evaluation benchmarks. Summarization, creative writing, open-ended QA, and value-pluralism benchmarks show negligible overlap ($\leq$ 1.6%). HumanEval, CRUXEval, IFEval, MATH, and WildBench show elevated overlap (7–30%), traceable to shared upstream sources: the Dolci SFT mixes include OpenThoughts3 (Guha et al., 2025), whose math questions derive from OpenMathInstruct-2 (Toshniwal et al., 2024), itself built on the MATH training set, large-scale Python corpora, Dolci-Think-Python (Olmo et al., 2025), Nemotron (NVIDIA, 2025) code split, and WildChat conversations (Zhao et al., 2024).

	EAD			SBERT			NLI			Vendi		
Task	T=1.0	T=0.6	Δ%	T=1.0	T=0.6	Δ%	T=1.0	T=0.6	Δ%	T=1.0	T=0.6	Δ%
TL;DR	0.478	0.365	-23.6	0.385	0.353	-8.1	0.987	0.949	-3.8	4.556	4.157	-8.8
CNN/DM	0.439	0.370	-15.8	0.254	0.279	+9.8	0.973	1.036	+6.5	2.928	3.162	+8.0
XSum	0.743	0.674	-9.3	0.551	0.451	-18.2	1.122	1.087	-3.1	6.628	5.192	-21.7
HumanEval	0.623	0.570	-8.5	0.438	0.411	-6.3	1.037	0.894	-13.7	2.578	2.454	-4.8
MBPP	0.631	0.592	-6.2	0.431	0.291	-32.5	1.236	1.179	-4.7	3.479	2.753	-20.9
CRUXEval	0.429	0.310	-27.6	0.293	0.239	-18.2	0.992	0.997	+0.6	2.841	2.458	-13.5
GSM8K	0.433	0.450	+3.9	0.199	0.172	-13.6	1.095	1.077	-1.7	2.251	2.094	-7.0
MATH-Algebra	0.472	0.453	-3.9	0.156	0.146	-6.2	1.012	1.000	-1.2	2.050	1.983	-3.2
MATH-Geometry	0.476	0.402	-15.5	0.210	0.198	-5.9	1.114	1.135	+1.8	2.525	2.438	-3.5
TruthfulQA	0.616	0.461	-25.2	0.452	0.353	-21.8	1.097	0.972	-11.4	5.173	3.805	-26.4
Alpaca	0.539	0.509	-5.6	0.396	0.319	-19.5	0.791	0.676	-14.5	2.620	2.217	-15.4
IFEval	0.554	0.443	-20.1	0.371	0.349	-6.0	1.063	1.055	-0.7	4.134	3.812	-7.8
WritingPrompts	0.357	0.234	-34.5	0.588	0.540	-8.1	1.166	1.165	-0.1	7.914	6.935	-12.4
PRISM	0.376	0.237	-37.0	0.452	0.408	-9.7	1.101	1.086	-1.4	5.313	4.639	-12.7
WildBench	0.475	0.300	-36.9	0.350	0.335	-4.3	1.087	1.064	-2.1	3.702	3.514	-5.1
Mean	0.509	0.425	-17.7	0.368	0.323	-11.2	1.058	1.025	-3.3	3.913	3.441	-10.3

Table 17: Base model diversity at recommended (T=1.0) vs. matched (T=0.6) temperature. Δ% reports the relative change.

	Think				Instruct				RL-Zero
Task	SFT	DPO	RL	Retain	SFT	DPO	RL	Retain	Retain
TL;DR	−53	−2	+1	46	−24	−19	+1	59	93
CNN/DM	−70	−8	+11	33	−20	−53	−1	26	71
XSum	−80	−6	+6	20	−37	−44	−0	18	94
HumanEval	−73	−7	+9	28	−73	−4	−0	23	105
MBPP	−72	+1	+1	31	−62	−13	−2	24	94
CRUXEval	−60	−8	+6	38	−9	−63	−3	26	103
GSM8K	−64	−6	−0	30	−39	+21	−37	45	95
MATH-Alg	−63	+5	+1	43	−10	−42	−9	39	95
MATH-Geo	−46	+3	+4	62	−9	−42	+2	51	92
IFEval	−45	−7	+9	56	−51	−5	+0	44	92
Alpaca	−56	−7	+9	46	−47	−15	−2	35	76
WritingPrompts	−56	−6	−1	37	−49	−9	−4	37	100
TruthfulQA	−66	−13	+0	21	−8	−48	−12	33	99
PRISM	−75	−3	+1	22	−65	−11	−1	22	95
WildBench	−52	−2	+6	52	−61	−2	−1	35	89
Average	**−62**	**−4**	**+4**	**38**	**−38**	**−23**	**−5**	**34**	**93**

Table 18: Per-task stage attribution: percentage of Base SBERT diversity lost (−) or recovered (+) at each post-training stage. *Retain* is the fraction of Base diversity preserved at the final checkpoint.

	CNN/DM	*XSum*	*TL;DR*	*GSM8K*	*MATH-Alg*	*MATH-Geo*	*HumanEval*	*MBPP*	*CRUXEval*	*Alpaca*	*IFEval*	*WrtPrompts*	*TruthfulQA*	*PRISM*	*WildBench*
Think-SFT	0.3	0.2	0.1	0.2	**10.2**	**15.4**	**21.5**	1.1	**9.5**	0.4	**7.6**	0.0	0.0	0.0	**20.5**
Think-DPO	0.2	0.3	0.0	0.0	1.2	1.2	**14.7**	0.0	**5.5**	0.7	**6.8**	0.0	0.0	0.0	**6.5**
Inst.-SFT	0.1	0.5	0.1	0.0	2.7	3.0	**30.1**	1.3	**11.2**	0.4	**7.3**	0.0	0.0	0.0	**26.4**
Inst.-DPO	0.1	0.1	0.0	0.0	1.8	1.8	**20.9**	0.3	**6.8**	1.6	**7.2**	0.0	0.0	0.0	**6.9**

Table 19: C_{13} 13-gram overlap (%) between Dolci training sets and evaluation benchmarks. Values $\geq 5\%$ are **bolded**.