

Reasoning–Answer Entailment Under GRPO and DPO Fine-Tuning

Hadi Mohammadi Anastasia Giachanou

Department of Methodology, Statistics & Data Science

Utrecht University, Utrecht, The Netherlands

{h.mohammadi, a.giachanou}@uu.nl

Abstract

Chain-of-thought (CoT) prompting can improve the problem-solving ability of large language models (LLMs), but a generated rationale does not always reflect the process by which a model reaches its answer, which limits the value of CoT for oversight and safety monitoring. We compare two post-training methods, Group Relative Policy Optimization (GRPO) and Direct Preference Optimization (DPO), on one measurable component of this problem: *reasoning–answer entailment*, whether the reasoning a model states supports the answer it gives. Entailment is a necessary condition for faithful reasoning, not a certificate of it, and we scope our claims accordingly. Across four scales of the Qwen2.5-Instruct family (1.5B–14B) on GSM8K, we evaluate eight fine-tuned models on accuracy, reliability, and entailment metrics computed from persisted per-example generations, report bootstrap confidence intervals and paired significance tests, and assess the robustness of LLM-as-a-judge scoring with two independent judges. GRPO yields significantly higher entailment at every scale (margins of +.14 to +.31, all $p < 10^{-7}$) and sweeps all metrics at 7B, while DPO matches GRPO’s greedy and pass@1 accuracy at 14B at lower training cost (self-consistency still favors GRPO). The methods fail differently. GRPO produces internally consistent rationales even when its answers are wrong; DPO’s stated reasoning fails to entail a majority of its correct answers at every scale, consistent with the DPO recipe optimizing answer style rather than the reasoning–answer link. A forensic analysis attributes the mid-scale (3B) collapse of both methods largely to output-format failure rather than lost reasoning ability, and a headline 14B judge-score gap does not survive re-evaluation with a second judge.

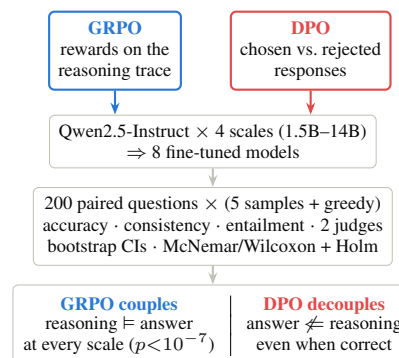


Figure 1: Overview: two post-training recipes, one scale-controlled evaluation with per-example statistics, and the central finding: GRPO’s stated reasoning entails its answers; DPO’s often does not.

1 Introduction

Chain-of-thought (CoT) reasoning (Wei et al., 2022) has become a standard technique for eliciting multi-step problem solving from large language models (LLMs). Beyond its effect on accuracy, the generated reasoning trace is increasingly treated as an interpretability artifact: a window into *how* the model reached its answer, and a channel through which humans or automated monitors can audit model behavior, mirroring the role explanations play across applied NLP more broadly (Mohammadi et al., 2025a; Mohammadi, 2026a). This use rests on an assumption of *faithfulness*: that the stated reasoning corresponds to the process that actually produced the answer (Jacovi and Goldberg, 2020).

That assumption is fragile. Models produce fluent but unfaithful justifications, change answers in response to hidden cues without acknowledging them, and construct persuasive rationales for conclusions reached by other means (Turpin et al., 2023; Lanham et al., 2023; Chen et al., 2025). When incorrect hints are embedded in XML meta-data, state-of-the-art reasoning models may shift

from correct to incorrect answers while presenting confident, coherent (and deceptive) reasoning chains (Chen et al., 2025). If CoT is to serve as an oversight mechanism, post-training should improve the fidelity of the reasoning along with its quality.

Most post-training pipelines, however, optimize for task performance or human preference, not reasoning fidelity. Two widely used methods embody contrasting philosophies. Direct Preference Optimization (DPO) (Rafailov et al., 2023) trains directly on pairs of preferred and rejected responses, dispensing with explicit reward models. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is an online reinforcement learning method that normalizes rewards within groups of sampled responses, and has been central to recent reasoning-focused models (Guo et al., 2025). Comparisons of preference-based and reinforcement-based post-training exist for general alignment (Xu et al., 2024; Iverson et al., 2024), but their comparative effect on the reasoning–answer link in CoT has not been systematically evaluated.

We address one measurable slice of this gap: a scale-controlled comparison of GRPO and DPO across four model scales (1.5B, 3B, 7B, 14B) of the Qwen2.5-Instruct family (Yang et al., 2024) on GSM8K (Cobbe et al., 2021), centered on *reasoning–answer entailment*: whether the reasoning a model states supports the answer it gives. We scope our claims deliberately. Entailment is necessary for faithful reasoning but does not certify it, and we run no causal interventions such as CoT truncation or corruption (Lanham et al., 2023); what we establish is which recipe produces reasoning that supports its answers, and how the two recipes fail when it does not. Figure 1 summarizes the design. Our contributions:

1. A scale-controlled comparison of GRPO and DPO on six metrics (accuracy, reliability, entailment, and judge scores), with bootstrap confidence intervals and paired significance tests computed from persisted per-example scores (§5);
2. A two-judge LLM-as-a-judge protocol that measures cross-judge agreement instead of assuming judge reliability (Thakur et al., 2025) (§5.3);
3. An error analysis over accuracy×entailment quadrants showing that the two methods fail

in qualitatively different ways: GRPO couples its answers to its stated reasoning (even when wrong), while DPO’s answers are frequently decoupled from the reasoning it states (§5.4);

4. Public release of code, per-example evaluation data, and all eight fine-tuned adapters.¹

Two further findings fell out of this protocol, and we report them as methodological notes rather than headline contributions: a forensic diagnosis of the performance collapse both methods exhibit at 3B, which we trace primarily to evaluation-time format failure using training-time reward trajectories as evidence (§5.2), and a second judge under which a prominent 14B judge-score gap does not replicate (§5.3).

2 Related Work

CoT faithfulness. Jacovi and Goldberg (2020) distinguish plausibility from faithfulness in NLP explanations and argue for evaluating the latter directly. For CoT specifically, Turpin et al. (2023) show that models rationalize biased answers without mentioning the bias; Lanham et al. (2023) propose causal probes (truncating or corrupting the CoT and observing answer changes) to quantify how much the stated reasoning matters; Chen et al. (2025) find that reasoning models verbalize hidden hints in only 1–20% of cases. Faithfulness tests based on input editing (Atanasova et al., 2023), self-consistency framings (Parcalabescu and Frank, 2024), and architectural guarantees (Lyu et al., 2023) offer complementary measurement strategies, while Bentham et al. (2024) caution that some unfaithfulness metrics track accuracy differences rather than genuine fidelity, and Arcuschin et al. (2026) document unfaithful CoT arising in the wild without adversarial prompting. Paul et al. (2024) move from measurement to intervention, training models so that answers causally depend on the produced reasoning. Our work sits beside this intervention-oriented line: we ask which of two standard post-training recipes yields reasoning that better entails its answers, holding data and scale constant.

Preference optimization and RL post-training. RLHF with PPO (Ouyang et al., 2022; Schulman

¹Code and per-example evaluation data: https://github.com/mohammadi-hadi/grpo_dpo_comparison. Fine-tuned adapters: <https://huggingface.co/hadimh93/qwen2.5-gsm8k-grpo-dpo-adapters>.

et al., 2017) established the reward-driven post-training paradigm; DPO (Rafailov et al., 2023) and successors (IPO, Gheshlaghi Azar et al., 2024; KTO, Ethayarajh et al., 2024; SimPO, Meng et al., 2024; ORPO, Hong et al., 2024) replace the RL loop with direct objectives over preference data. GRPO (Shao et al., 2024) retains online RL but replaces the value network with group-relative advantage estimates, powering DeepSeek-R1’s reasoning gains (Guo et al., 2025); RLOO (Ahmadian et al., 2024) makes a related simplification. Empirical comparisons across these families find that online RL methods tend to outperform DPO on general alignment (Xu et al., 2024; Ivison et al., 2024). The credit-assignment weakness of whole-response preference training for multi-step reasoning is documented: rejecting a full response penalizes its correct steps along with its errors, which motivated step-level variants such as Step-DPO (Lai et al., 2024) and Full-Step-DPO (Xu et al., 2025). We study the vanilla recipe deliberately, as the variant practitioners reach for first, and our results quantify how its weakness surfaces in the trained models: not as lower accuracy, but as answers that decouple from their stated reasoning, increasingly with scale. We extend the RL-versus-preference comparison axis to the reasoning–answer link, and use low-rank adaptation (LoRA)-based parameter-efficient training (Hu et al., 2022; Dettmers et al., 2023) to hold the fine-tuning budget constant across scales.

LLM-as-a-judge reliability. Judge models correlate with human preference at the system level (Zheng et al., 2023; Liu et al., 2023) but exhibit position and verbosity biases (Wang et al., 2024), self-preference (Panickssery et al., 2024), and leniency that varies with judge capability (Thakur et al., 2025); similar reliability concerns arise when LLMs are used as annotators, where model explanations affect annotation quality (Mohammadi et al., 2025b). Following recommendations from this literature, we score with two independent judges from different model families and report inter-judge agreement rather than treating any single judge as ground truth. Outcome-only judging has likewise been shown to miss failures that step-level inspection catches (Mohammadi, 2026b). Interpretable chain-of-thought has also been combined with LLM-as-judge peer review to evaluate LLM moral alignment (Mohammadi et al., 2026).

Consistency and statistical practice. Self-consistency (Wang et al., 2023) improves accuracy by majority-voting over sampled reasoning paths; we use agreement across samples as a reliability signal, and entailment-based scoring (Chen and Eger, 2023) as a lexical-overlap-free check that the reasoning supports the stated answer. Our statistical protocol (bootstrap intervals, paired exact tests, multiple-comparison correction) follows Dror et al. (2018) and Koehn (2004).

3 Method

3.1 Task and Data

We use GSM8K (Cobbe et al., 2021), a benchmark of 8.5K grade-school math word problems requiring multi-step reasoning, each with a worked solution and a numeric answer. For DPO we use a public GSM8K preference set² of $\sim 1.2\text{K}$ (prompt, chosen, rejected) triplets. To match training-set size, GRPO training samples $\sim 15\%$ of the GSM8K training split, of which 10% is held out for checkpoint selection. All evaluation uses a fixed set of 200 questions sampled from the GSM8K test split with a fixed seed; the same 200 questions are used for every model, enabling paired statistical tests. Appendix A shows both training formats on concrete examples.

3.2 Fine-tuning

We fine-tune Qwen2.5-Instruct models (Yang et al., 2024) at 1.5B, 3B, 7B, and 14B parameters with each method (eight models total), using QLoRA (quantized LoRA) (Hu et al., 2022; Dettmers et al., 2023): 4-bit quantized bases with LoRA adapters over all attention and MLP projections. We use this family because it provides instruction-tuned checkpoints at four sizes under one recipe and tokenizer, a clean scale ladder for paired comparison, and because much of the open GRPO reasoning work targeted it when these adapters were trained (see Limitations). Both methods use 8-bit AdamW at learning rate 5×10^{-6} and maximum sequence length 1024.

GRPO first adapts the model on ~ 80 examples to produce a structured output format (reasoning between `<start_working_out>`/`<end_working_out>` tags, answer between `<SOLUTION>`/`</SOLUTION>` tags). Training then samples 4 responses per prompt and scores

²<https://huggingface.co/datasets/valerielucro/gsm8k-preference-dataset-2.1>

them with four reward functions: exact format compliance, approximate format match, tiered answer correctness, and numerical-validity checks. LoRA rank 16, α 32. Compute constraints capped GRPO training at one epoch, with checkpoints saved at fixed step intervals and scored on the held-out split, which is disjoint from the 200 test questions; for each scale we deployed the highest-scoring saved checkpoint, which fell at 0.25 epochs (1.5B) and 0.5 epochs (3B–14B).

DPO trains on the preference triplets with the standard DPO objective ($\beta=0.1$), with its own output format (`<reasoning>/<answer>` tags). LoRA rank 32, α 64; three epochs with best-checkpoint selection on validation loss. Appendix B lists the full configuration.

Compute budget. Every run fits a single-GPU budget: 4-bit QLoRA bases (Dettmers et al., 2023), 8-bit AdamW (Dettmers et al., 2022), sequence length 1024, and a GRPO group of 4 samples per prompt (far smaller than the groups of 64 in Shao et al., 2024) keep even the 14B arms within one GPU’s memory. The scope is deliberate: we ask which recipe better couples reasoning to answers under a budget most practitioners actually have, not what either method achieves on a cluster. The asymmetries between the arms follow from the same budget. Each GRPO step samples and scores four full generations per prompt, so a GRPO epoch requires several times the generation compute of a DPO epoch; the GRPO runs therefore stopped within one epoch while DPO ran three.

Comparability caveats. The two arms are *representative recipes* rather than a controlled ablation: they differ in LoRA rank (16 vs. 32), epochs (fractional vs. 3), checkpoint selection, data format (raw problems with rewards vs. preference pairs), and output tags. Conclusions should be read as comparing the recipes as practitioners deploy them, not the objectives in isolation. Note the direction of the asymmetries: DPO trains adapters of twice the rank, sees its data for three full epochs, and selects its best checkpoint on validation loss, while GRPO trains fewer parameters for a fraction of an epoch. For the central entailment result the asymmetries run *against* the winner: the arm with fewer trained parameters and less training still shows higher entailment at every scale, so the gap cannot be attributed to a capacity or budget advantage for GRPO. The accuracy comparisons enjoy no such protection, and we flag where that matters (§5.1).

Whether the fractional-epoch GRPO checkpoints had converged cannot be settled from our data, but the deployed checkpoints demonstrably solve problems at training time (Figure 3), and the entailment margin recurs at all four scales, which an accident of early stopping would have to reproduce four times independently. Each method×scale cell is a single training run; our intervals capture variation over test questions, not training seeds (see Limitations).

3.3 Evaluation Framework

For each of the 200 test questions we draw $k = 5$ sampled generations (temperature 0.7, top- p 0.95, max 256 new tokens) plus one deterministic greedy generation, and persist every generation. Metrics:

- **Greedy accuracy:** correctness of the deterministic generation.
- **pass@1 (T=0.7):** correctness of the first sampled generation. (Our initial 4-bit protocol labeled this “greedy” accuracy; we keep the name for continuity and report true greedy decoding separately.)
- **Self-consistency accuracy** (Wang et al., 2023): correctness of the majority answer over the k samples.
- **Consistency ratio:** the fraction of parseable samples that agree with the majority answer.
- **Reasoning–answer entailment:** the entailment probability assigned by a natural language inference (NLI) model (RoBERTa-large-MNLI) to the hypothesis “The final answer is x .” given the generated reasoning as premise (Chen and Eger, 2023). We use this name rather than “NLI faithfulness”: the metric checks whether the stated reasoning supports the stated answer, a necessary but not sufficient condition for causal faithfulness (Jacovi and Goldberg, 2020; Parcalabescu and Frank, 2024).
- **LLM-as-a-judge:** a 1–5 rubric scoring of (problem, reasoning, answer, gold answer), normalized to $[0, 1]$, from two independent judges: Gemma-3-27B, which scored the initial 4-bit generations, and GPT-5-mini, which scored the bf16 re-evaluation. We report per-judge scores and their agreement (§5.3); the rubric is reproduced verbatim in Appendix C.

3.4 Statistical Protocol

All aggregate numbers come with 95% bootstrap confidence intervals (percentile method, 10,000 resamples over questions) (Koehn, 2004). GRPO–DPO differences at each scale are tested on paired per-question scores: McNemar’s exact test for binary metrics (greedy, pass@1, self-consistency) and Wilcoxon signed-rank for continuous ones (consistency ratio, entailment, judge scores), with Holm correction across the four scales within each metric (Dror et al., 2018).

4 Experimental Setup

The adapters were trained using 4-bit QLoRA and re-evaluated by attaching the same adapters to bf16 base models. All main results reported in this paper come from the bf16 re-evaluation, while the initial 4-bit results are used only for diagnostic comparison. Absolute scores should therefore be interpreted within the corresponding evaluation setting. Generation runs on Apple Silicon using the Metal Performance Shaders (MPS) backend, with the exact prompt templates from training (raw system/user/assistant pseudo-tags with each method’s XML-style tags). The NLI scorer runs locally; judge scoring uses the same Gemma-3-27B rubric verbatim for both judges. The re-evaluation persists per-example records for every generation. Compared with our initial 4-bit evaluation (3 samples per question, single judge, no per-example records and hence no paired tests), no comparison that reaches significance in the re-evaluation reverses the direction of the initial result, matching the initial “greedy” pass to pass@1, with two instructive exceptions analyzed below: 3B entailment (§5.2), which reverses, and the 14B judge gap (§5.3), which disappears.³

5 Results

5.1 Scaling Behavior

Table 1 and Figure 2 summarize performance across scales. The single most stable pattern is **entailment**: GRPO’s reasoning entails its answers significantly more often at *every* scale (1.5B: .539 vs. .319, Wilcoxon $p < 10^{-10}$; 3B: .427 vs. .176, $p < 10^{-11}$; 7B: .507 vs. .199, $p < 10^{-20}$; 14B: .491 vs. .353, $p < 10^{-7}$; all Holm-corrected). No other metric behaves this consistently.

³Initial 4-bit aggregate scores are included in the released data for comparison.

Accuracy tells a scale-dependent story. At 1.5B the methods are statistically indistinguishable on every accuracy metric (greedy: .470 vs. .425, McNemar $p = .37$; pass@1: $p = 1.0$; self-consistency: $p = .30$) and on judge scores ($p = 1.0$); the one non-entailment difference that survives correction favors DPO: more stable answers across samples (consistency ratio .664 vs. .528; $p < 10^{-5}$). At 7B the picture inverts sharply: GRPO wins *every* metric with large margins (greedy .750 vs. .430; self-consistency .875 vs. .610; judge .655 vs. .427; all $p < 10^{-6}$), including consistency ratio (.834 vs. .651, $p < 10^{-6}$), which had been DPO’s small-scale advantage. At 14B the gap narrows: GRPO retains significant advantages on self-consistency (.935 vs. .870, $p = .022$) and entailment, while greedy accuracy (.830 vs. .780, $p = .37$) and judge scores (.711 vs. .740, $p = 1.0$) are ties. GRPO’s advantage is thus not a uniform quality gap: it is concentrated on the reasoning–answer link, plus a transient mid-scale accuracy lead; §6 discusses the training-signal asymmetry consistent with this pattern.

True greedy decoding scores higher than pass@1 for seven of eight models (from +.04 to +.12; the exception is format-broken DPO-3B), confirming that the initial 4-bit protocol’s “greedy” label, actually a temperature-0.7 sample, understated deterministic performance and conflated two different quantities.

5.2 The 3B Collapse Is Largely a Format Failure

Both methods degrade sharply at 3B before recovering at 7B. Under 4-bit inference, GRPO-3B scored .055 on (pseudo-)greedy accuracy against .250 at 1.5B. Three independent sources of evidence indicate this collapse is dominated by evaluation-time output-format failure rather than lost reasoning ability.

First, GRPO training logs show the 3B model earned substantial answer-correctness reward *during training* (mean over the last ten logged steps up to the deployed checkpoint: 2.11, between the 1.5B model’s 1.52 and the 7B model’s 3.35; Figure 3): the checkpoint demonstrably solved problems while training. A capability collapse cannot explain a model that solves problems at train time and scores near zero at test time; a fragile output format can.

Second, per-example records expose how much scoring hinges on parseability. Already at 1.5B,

Size	Method	Greedy	pass@1	Self-Cons.	Cons. Ratio	Entailment	Judge
1.5B	DPO	.425 [.36, .50]	.345 [.28, .41]	.615 [.55, .69]	.664 [.62, .71]	.319 [.28, .36]	.379 [.32, .44]
	GRPO	.470 [.40, .54]	.350 [.29, .42]	.555 [.49, .63]	.528 [.49, .56]	.539 [.50, .58]	.375 [.31, .44]
3B	DPO	.135 [.09, .18]	.200 [.14, .26]	.510 [.44, .58]	.659 [.61, .71]	.176 [.14, .21]	.228 [.17, .28]
	GRPO	.265 [.20, .33]	.205 [.15, .26]	.530 [.46, .60]	.689 [.64, .74]	.427 [.38, .47]	.206 [.15, .26]
7B	DPO	.430 [.36, .50]	.390 [.33, .46]	.610 [.54, .68]	.651 [.59, .70]	.199 [.17, .23]	.427 [.36, .49]
	GRPO	.750 [.69, .81]	.650 [.58, .71]	.875 [.82, .92]	.834 [.80, .87]	.507 [.47, .55]	.655 [.59, .72]
14B	DPO	.780 [.72, .83]	.735 [.67, .80]	.870 [.82, .92]	.874 [.83, .91]	.353 [.32, .39]	.740 [.68, .80]
	GRPO	.830 [.78, .88]	.775 [.71, .83]	.935 [.90, .97]	.891 [.86, .92]	.491 [.45, .53]	.711 [.65, .77]

Table 1: Performance across methods and scales: mean [95% bootstrap CI] over 200 paired test questions. Bold marks the better method within a scale where the paired test is significant after Holm correction ($p < .05$); the Judge column reports GPT-5-mini scores.

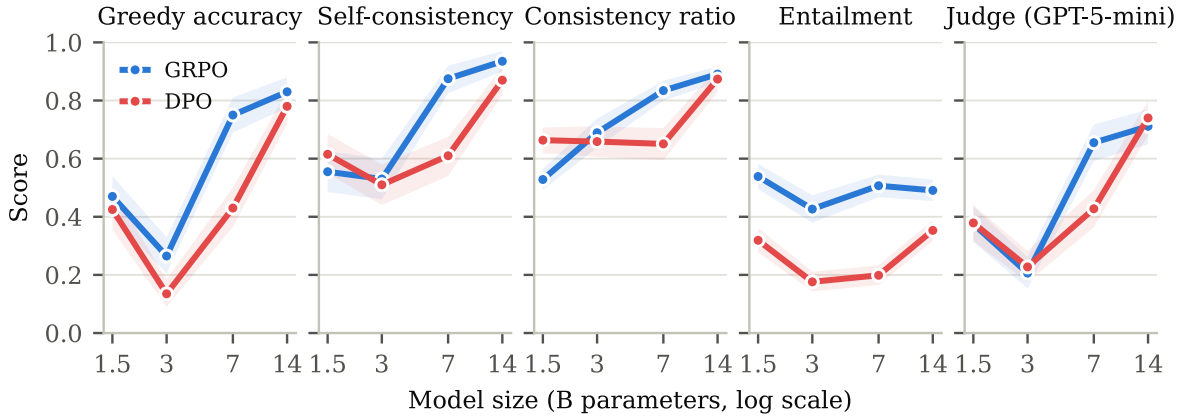


Figure 2: GRPO vs. DPO across model scale, with 95% bootstrap CI bands over the 200 paired test questions. pass@1 is omitted; it patterns with greedy accuracy (Table 1).

only 57% of DPO samples and 84% of GRPO samples contain a parseable answer in the required tag format; unparseable generations score zero on every accuracy metric regardless of the reasoning they contain. At 3B the parseable rate collapses to 37% (GRPO) and 34% (DPO), the worst of any scale, while at 7B it recovers to 82% and 52%. Conditioned on producing at least one parseable answer, 3B self-consistency is .61 (GRPO) and .63 (DPO), close to the corresponding 1.5B values (.56 and .68) rather than collapsed: the models reason at roughly their 1.5B level and format catastrophically. The 256-token generation cap compounds the effect: reasoning that overruns the cap never emits its closing tags. Per-model compliance rates are listed in Appendix E.

Third, the collapse largely lifts under full-precision (bf16) re-evaluation. Under 4-bit inference GRPO-3B scored .055 (pseudo-)greedy accuracy and .120 self-consistency; our bf16 re-evaluation of the *same adapter* on the same questions reaches .265 and .530 respectively. The sub-

stantial recovery under bf16 evaluation suggests that precision-related format instability contributed to the 3B degradation. This also explains why the 3B comparisons change direction between the two evaluations, most consequentially entailment, which the 4-bit evaluation’s format failures had driven to near-zero for GRPO and which flips to a significant GRPO advantage (.427 vs. .176, $p < 10^{-11}$) once formatting recovers.

This reframes the intuitive interpretation (“hyperparameter sensitivity at a capacity threshold”): uniform hyperparameters did hurt the 3B models, but primarily by destabilizing their *output formatting*, and metrics that gate on format then amplified the damage. Evaluations of tag-constrained models should report format-compliance rates alongside accuracy.

5.3 Judge Agreement

Replacing the initial Gemma-3-27B judge with GPT-5-mini shifts absolute scores substantially (at 1.5B, GRPO’s judge score rises from .120 to .375),

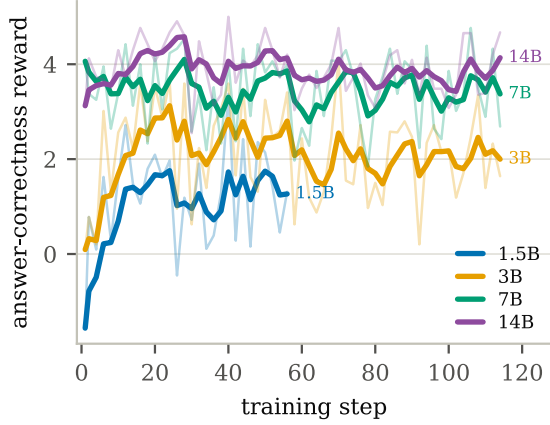


Figure 3: GRPO per-step answer-correctness reward by model size (bold: exponentially smoothed; faint: raw logged steps; from released trainer states). The 3B run earns intermediate reward, inconsistent with a training-time capability collapse.

underlining that a single judge’s absolute scale is not meaningful across studies (Thakur et al., 2025). Across the eight systems, the two judges agree at the coarse level (Kendall $\tau = .71$, $p = .014$; Spearman $\rho = .83$; Figure 4): both rank the 14B models on top and Gemma is systematically harsher on small-model outputs. Because Gemma scored the initial 4-bit generations and GPT-5-mini the bf16 ones, this comparison changes the judge and the evaluation setting together; the system-level ranking is what the two runs share.

Where the judges *disagree* is instructive. A headline judge result, GRPO-14B beating DPO-14B by $+.172$ (.748 vs. .576) under Gemma, does not replicate under GPT-5-mini, which scores the pair as a statistical tie (.711 vs. .740, $p = 1.0$). A judge-score gap between two strong systems can thus fail to survive a change of judge. System-level conclusions in this paper rest on judge scores only where both judges agree; we treat LLM-judge output as a ranking signal, not a calibrated quantity, and never as a faithfulness measure on its own.

Two further limits of this protocol deserve naming. Both judges apply the same five-level rubric (Appendix C), and its adjacent levels separate on wording that a human rater might also find hard to apply consistently; the coarse, system-level agreement we observe is consistent with that. More fundamentally, agreement between two judges bounds reliability but cannot arbitrate disagreement: when the judges diverge, nothing in our protocol says which one is closer to the truth, and neither judge,

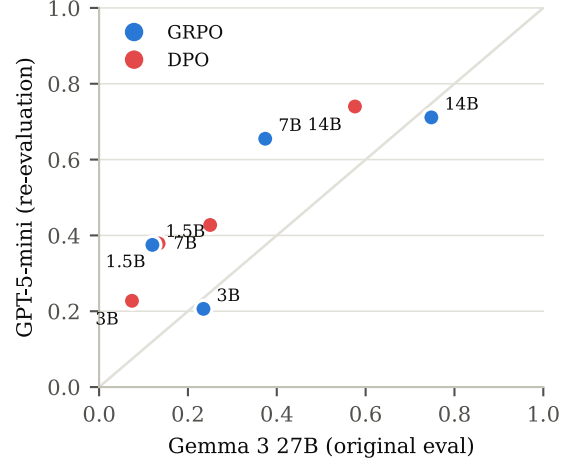


Figure 4: System-level judge agreement: each point is one fine-tuned model, scored by Gemma-3-27B on the initial 4-bit generations and by GPT-5-mini on the bf16 re-evaluation. Judges agree on coarse ranking ($\tau = .71$) but not on the 14B GRPO–DPO gap.

nor the NLI scorer, has been calibrated against human judgments on this data. A small human-annotated anchor set, or a third judge for majority adjudication, would turn the protocol from a robustness check into a validation; it is the clearest next step, and we have not taken it.

5.4 Accuracy $\times$ Entailment Quadrants

Crossing self-consistency correctness with entailment (threshold .5) splits the 200 questions into four outcomes (Figure 5). At 1.5B the two methods occupy strikingly different regions. GRPO’s errors are dominated by *internally consistent* mistakes: 28.0% of its questions land in the wrong-but-entailed quadrant (the model states reasoning that genuinely supports its incorrect answer) versus only 7.0% for DPO. DPO instead concentrates in correct-but-not-entailed: 35.0% of its questions are answered correctly with stated reasoning that does *not* entail the answer (GRPO: 24.0%), and its wrong answers overwhelmingly come with non-entailing reasoning too (31.5% wrong-and-not-entailed vs. GRPO’s 16.5%).

The interpretation cuts both ways, and we resist declaring a winner from this quadrant alone. GRPO’s profile is what one expects from a model whose answers are *coupled* to its stated reasoning: when the reasoning goes wrong, the answer follows it, a transparent and auditable failure. DPO’s profile is the signature of *decoupling*: answers (right or wrong) are produced with reasoning that does not

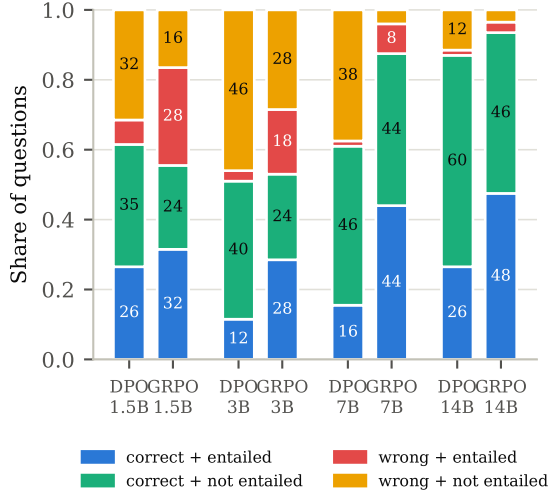


Figure 5: Composition of correct/wrong $\times$ entailed/not-entailed outcomes per model (share of 200 questions; direct labels are percentages).

support them, consistent with preference-pair training teaching answer style rather than reasoning–answer dependence. For oversight purposes, the DPO pattern is arguably the more concerning one (a monitor reading DPO’s reasoning learns little about where its answers come from), while GRPO’s wrong-but-entailed cases are the classic “fluent-but-wrong” risk: persuasive rationales for incorrect conclusions.

Scale changes the magnitudes but not the asymmetry. GRPO’s wrong-but-entailed share shrinks steadily as accuracy improves (28.0% $\rightarrow$ 18.5% $\rightarrow$ 8.5% $\rightarrow$ 3.0% from 1.5B to 14B), so the fluent-but-wrong risk is predominantly a small-model phenomenon. DPO’s decoupling, by contrast, does not fade with scale: a majority of its correct answers go unentailed at every scale (57–78%, against GRPO’s 43–50%), and at 14B, 60.5% of its questions are answered correctly with stated reasoning that does not entail the answer (GRPO-14B: 46.0%). Appendix D shows a representative pair: DPO lists correct intermediate steps but never completes the derivation, so its (correct) answer is asserted rather than derived.

5.5 Metric Correlations

Correlating accuracy metrics over eight model-level aggregates ($n = 8$) would suggest $r > .85$, but $n = 8$ cannot support such a claim. Example-level correlations over all per-question data ($n = 1,600$ method–question pairs) paint a more informative picture: the accuracy family does cohere

(self-consistency $\times$ pass@1: Spearman $\rho = .62$; $\times$ consistency ratio: $\rho = .56$; $\times$ judge: $\rho = .54$; all $p < 10^{-100}$), but entailment correlates only weakly with all of it (entailment $\times$ self-consistency: $\rho = .33$; $\times$ consistency ratio: $\rho = .26$; $\times$ judge: $\rho = .24$). Entailment and judge scores, both called “faithfulness” metrics in prior work, share under 6% of rank variance: they measure different constructs, and studies should not treat them as interchangeable.

6 Discussion

What the comparison actually shows. Once per-example uncertainty is accounted for, the GRPO–DPO contrast is not a uniform quality gap but a dissociation along the reasoning–answer link. GRPO (trained by sampling full reasoning traces and rewarding their format along with the correctness and numerical validity of the answers they reach) produces rationales that entail its answers, significantly at every scale, with margins between $+.14$ and $+.31$. DPO (trained to prefer whole responses) produces stable, well-styled answers whose stated reasoning frequently fails to support them, and this decoupling persists at every scale, including where accuracy converges. The mid-scale accuracy sweep (GRPO wins everything at 7B) narrows to parity at 14B, where a single-judge quality gap turns out not to replicate across judges: what survives at the largest scale is precisely the entailment advantage, consistent with GRPO’s training signal being computed on sampled reasoning traces while DPO’s never is.

Practical guidance. For deployments where the reasoning trace will be *read* (audit pipelines, tutoring, any oversight setting), GRPO-style reward shaping on the trace is worth its extra compute and tuning fragility: of the two recipes it is the one that yields the stronger reasoning–answer link at every scale, in the entailment sense this paper measures (see Limitations). Where only answer accuracy matters and compute is constrained, DPO is competitive at 14B (all accuracy metrics tie except self-consistency) at lower training cost, since GRPO samples four responses per prompt online whereas DPO trains offline on fixed preference pairs. The 3B forensics adds an operational lesson: tag-constrained evaluation without format-compliance reporting can misdiagnose a formatting regression as a reasoning collapse.

Metrics for faithfulness. Neither entailment nor judge scores certify causal faithfulness (Lanham et al., 2023; Parcalabescu and Frank, 2024), and our correlation analysis shows they are largely independent signals. Reporting both, with uncertainty, alongside intervention-based probes where affordable, is the defensible protocol; single-metric, single-judge faithfulness claims are fragile in both directions.

7 Conclusion

We compared GRPO and DPO by their effect on reasoning–answer entailment in chain-of-thought outputs, across four scales of Qwen2.5-Instruct, with per-example statistical testing, two-judge agreement analysis, and failure-mode forensics. GRPO yields a significantly stronger link between stated reasoning and answers than DPO at every scale (entailment .539 vs. .319 at 1.5B, $p < 10^{-10}$; .491 vs. .353 at 14B, $p < 10^{-7}$) and sweeps all metrics at 7B; DPO matches GRPO’s greedy and pass@1 accuracy at 14B but leaves its answers decoupled from the reasoning it states. The 3B collapse exhibited by both methods is largely attributable to output-format failure rather than a loss of reasoning capability, and a headline 14B judge gap does not survive re-evaluation with a second judge. Future work includes adaptive hyperparameters for small scales, causal faithfulness probes during training, a human-annotated anchor set to calibrate the automatic scorers, reward terms that explicitly penalize reasoning–answer decoupling, and replication beyond this model family and beyond mathematical reasoning.

Limitations

Our comparison covers one model family (Qwen2.5-Instruct) and one domain (grade-school math); conclusions may not transfer to other architectures or reasoning tasks. We chose the family for the reasons given in §3.2, and it has since been superseded within its own line (Yang et al., 2025); cross-family replication is the most direct extension of this work. The two training arms differ in LoRA rank, epochs, checkpoint selection, and data format (§3.2); we frame the comparison at the recipe level and report all settings, but objective-level causal attribution would require a full ablation grid. Our entailment and judge metrics measure whether stated reasoning supports the stated answer and how plausible it appears, necessary but not sufficient condi-

tions for causal faithfulness (Jacovi and Goldberg, 2020); we ran no intervention-based probes such as CoT truncation (Lanham et al., 2023). The entailment quadrants use a fixed .5 threshold; shares shift with the threshold although the GRPO–DPO asymmetry does not. Training used 4-bit QLoRA, whereas the reported re-evaluation used bf16 base models, which should be considered when comparing absolute scores with the initial 4-bit evaluation. Each method is represented by a single training run per scale, so our confidence intervals and p -values capture variation across the 200 test questions, not across training seeds; seed-level replication would be needed to generalize fully to the methods rather than to these specific checkpoints. We also report no untuned base-model baseline, so absolute lift over vanilla Qwen2.5-Instruct is out of scope. Finally, evaluation uses 200 test questions with $k = 5$ samples; while sufficient for the paired tests we report, larger samples would tighten the intervals.

No human evaluation backs the automatic metrics. The NLI scorer was not validated on this dataset, and neither judge was calibrated against human raters; the two-judge protocol detects disagreement but cannot resolve it (§5.3). Findings that rest on judge scores are therefore weaker than those that rest on accuracy or entailment, and we use them only where both judges agree.

We also ran no contamination audit. GSM8K predates the Qwen2.5 pretraining corpus and is among the most heavily scraped benchmarks; overfitting to it is documented (Zhang et al., 2024). Contamination would inflate absolute accuracy, and it bears on entailment too: a model that can retrieve an answer without deriving it only needs to narrate toward that answer, so entailment would partly measure narration quality rather than derivation. Two features of the design limit the damage to our comparative claims. Both arms fine-tune the same base and are evaluated on the same 200 items, so any contamination is matched across the comparison, and the entailment gap persists on questions the models answer incorrectly (§5.4), where mean entailment is .54 versus .18 at 1.5B and .43 versus .10 at 14B and a memorized answer cannot be supplying the result. Absolute scores should nevertheless be read with this caveat.

Ethics Statement

This work aims to improve the transparency and auditability of LLM reasoning. It uses only pub-

licly available datasets of math word problems; no human subjects or personal data are involved. Post-training that strengthens the reasoning–answer link can reduce the risk of models presenting rationales disconnected from their answers in high-stakes applications. Two caveats temper this: entailment-style metrics can be optimized to make reasoning *appear* faithful without guaranteeing causal faithfulness, and training the eight models carries a compute and energy cost.

Data Availability

GSM8K and the DPO preference set are publicly available at their original sources. We release the complete evaluation pipeline, per-example records for all 9,600 generations scored in this paper (8 models $\times$ 200 questions $\times$ 6 generations), aggregate result tables, GRPO trainer states, and all eight LoRA adapters: code and data at https://github.com/mohammadi-hadi/grpo_dpo_comparison, adapters at <https://huggingface.co/hadimh93/qwen2.5-gsm8k-grpo-dpo-adapters>.

Acknowledgments

We thank Tamas Kozak for training the eight adapters analyzed in this paper. We also thank the three anonymous INLG reviewers and the area chair, whose reports pushed us to scope the paper’s central claim to reasoning–answer entailment and to state more precisely what the comparison does and does not control.

References

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand. Association for Computational Linguistics.
- Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2026. [Chain-of-thought reasoning in the wild is not always faithful](#). In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR.
- Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- Oliver Bentham, Nathan Stringham, and Ana Marasović. 2024. [Chain-of-thought unfaithfulness as disguised accuracy](#). *Transactions on Machine Learning Research*.
- Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. [Reasoning models don’t always say what they think](#). *Preprint*, arXiv:2505.05410.
- Yanran Chen and Steffen Eger. 2023. [MENLI: Robust evaluation metrics from natural language inference](#). *Transactions of the Association for Computational Linguistics*, 11:804–825.
- 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. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. [8-bit optimizers via block-wise quantization](#). In *International Conference on Learning Representations*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [Model alignment as prospect theoretic optimization](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12634–12651. PMLR.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. [A general theoretical paradigm to understand learning from human preferences](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang,

- Shirong Ma, Xiao Bi, et al. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. [ORPO: Monolithic preference optimization without reference model](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. [Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. [Step-DPO: Step-wise preference optimization for long-chain reasoning of LLMs](#). *Preprint*, arXiv:2406.18629.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukošiuūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *Preprint*, arXiv:2307.13702.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful chain-of-thought reasoning](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. [SimPO: Simple preference optimization with a reference-free reward](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Hadi Mohammadi. 2026a. *Let Me Explain!: Explainable NLP for Understanding Large Language Models*. Ph.D. thesis, Utrecht University.
- Hadi Mohammadi. 2026b. [trajectory-judge: What outcome-only LLM judges miss on agent trajectories](#). *Preprint*, arXiv:2609.00038.
- Hadi Mohammadi, Robert A. Bagheri, Anastasia Giachanou, and Daniel L. Oberski. 2025a. [Explainability in practice: A survey of explainable NLP across various domains](#). *Preprint*, arXiv:2502.00837.
- Hadi Mohammadi, Anastasia Giachanou, and Robert A. Bagheri. 2026. [EvalMORAAL: Interpretable chain-of-thought and LLM-as-judge evaluation for moral alignment in large language models](#). In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 497–515, San Diego, California, United States. Association for Computational Linguistics.
- Hadi Mohammadi, Tina Shahedi, Pablo Mosteiro, Massimo Poesio, Ayoub Bagheri, and Anastasia Giachanou. 2025b. [Assessing the reliability of LLMs annotations in the context of demographic bias and model explanation](#). In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 92–104, Vienna, Austria. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Letitia Parcalabescu and Anette Frank. 2024. [On measuring faithfulness or self-consistency of natural language explanations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

- Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. [Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. [Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35.
- Huimin Xu, Xin Mao, Feng-Lin Li, Xiaobao Wu, Wang Chen, Wei Zhang, and Anh Tuan Luu. 2025. [Full-Step-DPO: Self-supervised preference optimization with step-wise rewards for mathematical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24343–24356, Vienna, Austria. Association for Computational Linguistics.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. [Is DPO superior to PPO for LLM alignment? A comprehensive study](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54983–54998. PMLR.
- An Yang, Anfeng Li, Baosong Yang, et al. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024. [A careful examination of large language model performance on grade school arithmetic](#). In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems*, volume 36. Datasets and Benchmarks Track.

A Dataset and Format Examples

GRPO training format. *Question:* “Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?”

```
<start_working_out>
Natalia sold 48/2=24 clips in May.
Total: 48+24=72 clips altogether.
<end_working_out>
<SOLUTION>72</SOLUTION>
```

DPO preference pair. *Prompt:* “Mr. Sam shared money between Ken (\$1750) and Tony (twice Ken’s amount). Total shared?” *Rejected:* claims “insufficient information” despite having all data. *Chosen:* correctly calculates $2 \times \$1750 = \$3500$, total = \$5250.

	GRPO	DPO
LoRA rank / α	16 / 32	32 / 64
Learning rate	5×10^{-6}	5×10^{-6}
Epochs	≤ 1 (ckpt. 0.25–0.5)	3 (best val.)
Batch generations	4 per prompt	n/a
β	n/a	0.1
Max seq. length	1024	1024
Optimizer	AdamW 8-bit	AdamW 8-bit
Precision	4-bit QLoRA	4-bit QLoRA

Table 2: Training hyperparameters per method.

B Hyperparameters

Models were trained with Unsloth 4-bit QLoRA on CUDA GPUs; per-run trainer states are in the released logs. The GPT-5-mini judge model version and both judges’ decoding settings are recorded with the released code.

C Prompts and Judge Rubric

Evaluation prompts. Each model is prompted with the same raw pseudo-tag template used in its training: a system message specifying the tag format, the question as user message, and an assistant prefix opening the reasoning tag.

Judge rubric (identical for both judges):

Evaluate this mathematical solution on a scale of 1–5: [problem, student reasoning, student answer, correct answer]

1 = Poor reasoning AND wrong answer; 2 = Some logical steps but significant issues OR good reasoning with wrong answer; 3 = Generally sound reasoning with minor issues, answer may be wrong due to arithmetic errors; 4 = Strong reasoning with very minor flaws, correct or nearly correct answer; 5 = Excellent reasoning with correct answer.

Respond with only a number from 1 to 5.

D Qualitative Example

One sampled generation per method for the same 1.5B test question, lightly cleaned from the released per-example data (calculator annotations removed, elisions marked), illustrating the coupling contrast of §5.4 at the level of single generations. *Question:* Stephen’s grocery order came to \$40.00; the delivery vendor adds a 25% fee and a \$3.00 delivery charge, and Stephen adds a \$4.00 tip. Gold answer: 57.

GRPO (wrong but entailed; entailment .85).

The model makes one modeling error (applying the 25% fee after adding the delivery charge and tip) and then carries its own reasoning through to the answer it states:

The grocery bill itself is \$40.00. The delivery fee was \$3.00 so that makes \$40 + \$3 = \$43. The tip was \$4.00 so now we have \$47. But there was a 25% service charge on top of that which means we need to calculate $.25 \times 47 = \$11.75$. So our final cost is $\$47 + \$11.75 = \$58.75 \dots$ [answer: 58.75]

DPO (correct but not entailed; entailment

.002). The model lists all the correct components but never completes the derivation; the correct answer is asserted, not derived from the stated reasoning:

1. The delivery vendor adds a 25% fee. . . $25/100 \times \$40.00 = \$10.00$. 2. In addition. . . there were \$3.00. . . the new total before adding the tip would be $\$40.00 + \$3.00 + \$10.00$. 3. Finally, we add the \$4.00 tip. . . 4. Let’s calculate the total after all these additional charges: $\$40.00 + \$3.00 + \$10.00 + \$4.00$. [reasoning ends; answer: 57.00]

The NLI scorer cannot entail “The final answer is 57” from reasoning that stops before summing, an honest reflection of the decoupling pattern: DPO’s answers appear *alongside* supporting computation rather than as its conclusion.

E Additional Results

Format-compliance (parseable-answer) rates per model, sampled generations / greedy pass: GRPO .84/.90 (1.5B), .37/.40 (3B), .82/.84 (7B), .86/.95 (14B); DPO .57/.60 (1.5B), .34/.28 (3B), .52/.58 (7B), .75/.80 (14B). GRPO’s format rewards yield higher compliance at every scale; both methods crater at 3B (§5.2). Full per-example records, the complete significance table, and the initial 4-bit aggregates are in the released data.