BOOTSTRAPPING POST-TRAINING SIGNALS FOR OPEN-ENDED TASKS VIA RUBRIC-BASED SELF-PLAY ON PRE-TRAINING TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-play has recently emerged as a promising paradigm for post-training Large Language Models (LLMs). In self-play, the target LLM creates the task input (e.g., a question), which it then addresses itself by producing a task output (e.g., an answer). A reward model evaluates the output, and the rewards are used to train the LLM, typically via Reinforcement Learning (RL). A key benefit of self-play for post-training LLMs is its minimal supervision costs: self-play avoids the need for high-quality input-output pairs traditionally constructed by humans or expensive proprietary models. Existing work, however, explores self-play only for verifiable tasks, such as math and coding, for which objective ground truth is available and easily checkable. In this paper, we seek to extend self-play to more realistic open-ended tasks. We propose POP, the *first fully autonomous LLM self-play framework* on open-ended tasks. It uses the same LLM to synthesize evaluation rubrics along with each input-output pair. The rubric is used to evaluate outputs and train the model. Crucially, we ground the framework on a content-rich pretraining corpus to (1) enable an exploitable generation-verification gap and reduce reward hacking, and (2) prevent mode collapse. On Qwen-2.5-7B, POP increases the performance of both the pretrained base model and instruction-tuned model on multiple tasks ranging from long-form healthcare QA to creative writing and instruction following.

1 INTRODUCTION

Recent Large Language Models (LLMs) have become increasingly capable of handling complex tasks, from math problem solving (Luong & Lockhart, 2025) to agentic workflows (Plaat et al., 2025). However, continual post-training of LLMs requires high-quality data from humans or stronger LLMs, which is often expensive or unavailable. In fact, the cost and scarcity of high-quality training data is becoming a bottleneck (Villalobos et al., 2024).

To this end, self-play has emerged as a promising direction for post-training LLMs with minimal external supervision. In essence, in each iteration of self-play (i) the LLM synthesizes task input (e.g., asks a question); (ii) the same LLM responds with task output (e.g., gives an answer); (iii) the output (and optionally the input) is scored by a task-dependent reward model (e.g, a rule-based verifier; neural network, etc.); and (iv) the reward is used to update the LLM, to make it produce both better task-specific output (e.g., answer better) and (optionally) better task input (e.g., ask better questions). In contrast to standard RL and alternative self-improvement methods (Huang et al., 2023; Yuan et al., 2024), self-play avoids the need for supervision not only of the task output, but also the task input.

Self-play for post-training has thus far only been explored in "verifiable" domains such as math Huang et al. (2025a); Liu et al. (2025); Zhang et al. (2025); Chen et al. (2025) and coding Zhao et al. (2025); Zhou et al. (2025b), where verifying the correctness of task output is easy. Extending self-play to more realistic open-ended tasks is important, but challenging. Recent work Gunjal et al. (2025); Viswanathan et al. (2025); Bi et al. (2025); Zhou et al. (2025a); He et al. (2025); Rezaei et al. (2025); Huang et al. (2025b); Shao et al. (2025), however, shows that rubric-based rewards can provide reliable signals for open-ended tasks. Inspired by this, we propose a rubric-based self-play framework that post-trains LLMs for open domains. In particular, after the model generates responses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

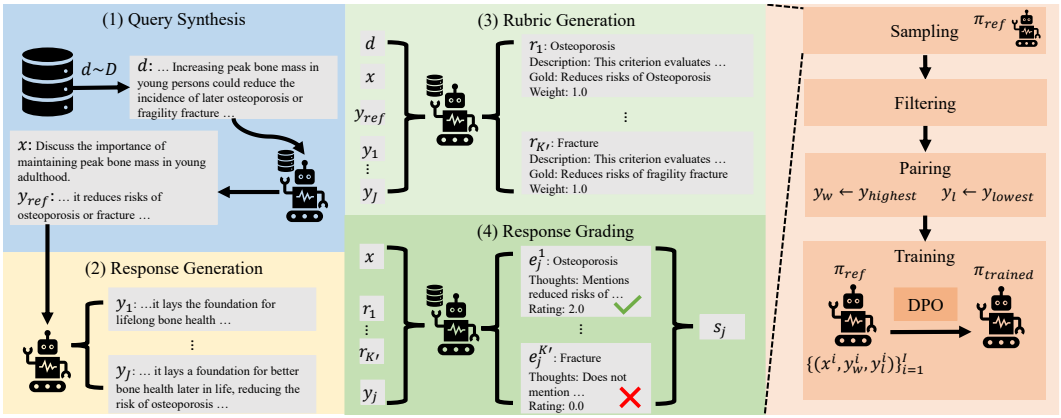


Figure 1: Training pipeline for POP. We sample a set of examples from the base model and create response pairs from them, which are then used to train the model via DPO. To create each example, (1) we sample a piece of text from a pre-training corpus and, grounded on that, the LLM synthesizes a query and a reference response; (2) the LLM produces a set of candidate responses to the query; (3) the LLM generates an evaluation rubric, conditioned on the pre-training text; (4) the LLM scores each response according to the rubric. After sampling, we filter out invalid examples and take the responses with the highest and lowest scores as the preferred and dispreferred responses, respectively.

to a query, we use the *same* model to generate an evaluation rubric specific to that query, and then grade each response based on the rubric.

The use of the same model to synthesize queries, responses, rubrics, and rewards engenders several risks: (1) limited query diversity, causing mode collapse; (2) poor response quality, limiting positive training signals and learning effectiveness; (3) an insufficient generation-verification gap Song et al. (2025) (i.e., ensuring that response verification is easier than response generation), resulting in reward hacking. We address (1) by grounding the framework on diverse text sampled from a content-rich pre-training corpus. We ask the model to synthesize queries that draw from the text, and, in tasks for which ground truth exists, to construct queries whose answers are derivable from the text. We address (2) by using Direct Preference Optimization (DPO) Rafailov et al. (2023), to train the LLM via contrastive signals from the response pairs. We use two methods to mitigate (3). First, we give privileged access to the pre-training text to the model when it evaluates its generated responses. Second, we select only the highest and lowest scored responses for DPO training.

Our resulting approach **P**ost-trains LLMs for **O**pen domains with **P**retraining text (**POP**). POP uses the same model in three roles: the Proposer, which synthesizes queries, grounded on pre-training text; the Solver, which answers the queries; and the Verifier, which generates the rubric and grades each Solver's response, again grounded on pre-training text. We evaluate POP on three open-ended tasks: long-form healthcare QA, creative writing, and general instruction following. POP is generalizable across domains and has *minimal priors specific to any testing benchmarks*: For each task, we change only the pre-training corpus and the prompts for the proposer and solver, while keeping others fixed.

Our main contributions are as follows. (i) We propose POP, *the first fully autonomous self-play framework* for post-training LLMs on open-ended tasks; (ii) We propose the use of pre-training text to ground POP to enable an exploitable generation-verification gap without introducing strong supervision; (iii) Experiment results show that POP increases performance of pretrained base models and instruction-tuned models on three quite distinct tasks — long-form Healthcare QA (up to 3% on HealthBench500 Arora et al. (2025)), creative writing (up to 3% on Creative Writing V3 Paech (2025)), and instruction following (up to 5% on IFEval Zhou et al. (2023) and 4% on ArenaHard Zheng et al. (2023)); (iv) Further analysis suggests that while our rubric evaluations are still imperfect, pairing the highest and lowest scored responses produces mostly correct training signals.

2 Related Work

Self-Play. Prior work applies self-play almost exclusively to verifiable reasoning tasks: Zhao et al. (2025) and Zhou et al. (2025a) apply self-play to coding tasks; Chen et al. (2025); Huang et al.

(2025a); Liu et al. (2025); Zhang et al. (2025) focus on math reasoning; Cheng et al. (2024) works with verifiable commonsense reasoning. DeepSeek-AI (2025) and Research (2026) synthesize question-answer pairs on various verifiable agentic tasks, with access to tools that interact with environments. However, these verifiable tasks are only a small part of what is needed to train highly capable LLMs (Chung et al., 2022). Real-world use cases often require long-form open-ended output, such as document-assisted writing (e.g., summarization), creative writing (e.g., storytelling), information extraction, open-ended QA, etc. POP focuses on self-play for open-ended tasks.

Rubric as Rewards. There is a growing body of literature on extending existing Reinforcement Learning (RL) methods to post-train LLMs in non-verifiable domains. One major branch uses rubrics to score open-ended text from the model, and uses the score as the reward for RL. Without exception, these methods require access to humans or strong LLMs to generate rubric or evaluate answers, both of which are expensive and, in some cases, unavailable.

Specifically, Dineen et al. (2025) rely on strong LLMs to synthesize hierarchical rubrics according to given principles. Huang et al. (2025b) synthesize the rubrics in advance, from human or strong LLMs, and use these rubrics to synthesize questions for training. Gunjal et al. (2025); Viswanathan et al. (2025) use strong LLMs to generate question-specific rubrics. Rezaei et al. (2025) ask strong LLMs to compare responses from the current model against those from a reference model to curate new rubric criteria during training. He et al. (2025) use human-annotated rubrics and scores to train their LLM-based rubric generators and verifiers. Zhou et al. (2025a); Bi et al. (2025) use human or strong-LLM-synthesized rubrics to guide the solver to generate high-quality responses, to either enable curriculum learning or boost training efficiency. In contrast to all of the above, POP does not rely on strong supervision for rubric evaluation.

3 METHODOLOGY

The full training pipeline of POP is shown in Figure 1; for pseudo code, see Algorithm 1. POP is a rubric-based self-play framework that uses the *same* model to (1) synthesize queries, (2) generate responses, (3) generate rubrics, and (4) grade the responses. Steps (1), (3), (4) are grounded on the pre-training text. The rubric scores are used to construct response pairs for DPO training.

3.1 SAMPLING

We presume access to a pre-training corpus D with content relevant to the target task t (e.g., a corpus of medical articles or clinical records for health QA; fantasy books for creative writing; general text for instruction following). Given D, we prompt a base LLM π_{ref} to synthesize I examples of t. Each example is synthesized according to the following four steps.

Query Synthesis (Proposer). First, a document d is sampled from D. To ensure that d is informative, we force it to be at least 50 words long. For efficiency reasons, we also truncate d to the first 1024 words. Next, the LLM is prompted to synthesize a new query x and a reference response y_{ref} conditioned on d: $(x, y_{ref}) \sim \pi_{ref}(\cdot|P_t^{proposer}(d))$, where $P_t^{proposer}$ is a task-specific query synthesis prompt (see Appendix R for all prompts). y_{ref} will be used later for rubric generation.

For tasks where ground truth answers exist (e.g., long-form health/medical QA), we ask in the prompt that the answer to x be derivable from d. For tasks where ground truth does not exist (e.g., creative writing; instruction-following), we relax the condition and ask that the query be drawn from d.

Response Generation (Solver). A set of candidate responses to x is sampled from π_{ref} according to a task-specific response-generation prompt P_t^{solver}: $\forall j \in \{1, \cdots, J\}, y_j \sim \pi_{ref}(\cdot|P_t^{ans}(x))$.

Rubric Generation. A *query*-specific rubric r is generated by π_{ref} according to a rubric generation prompt P^{rubric} and conditioned on the pre-training document d, query x, reference response y_{ref}, and candidate responses $\{y_j\}_{j=1}^J$: $r \sim \pi_{ref}(\cdot|P^{rubric}(d, x, y_{ref}, \{y_j\}_{j=1}^J))$.

r consists of K' criteria $r_1, \cdots, r_{K'}$, where K' is upper bounded by a constant K. Each criterion r_k has a name, a description of the characteristics of good and bad responses, an optional gold label g_k extracted from the document d, and a weight w_k that indicates how important the criterion is.

Importantly, we enforce several principles via the rubric generation prompt. First, the rubric must be **grounded on the pre-training document** d. For criteria that have ground truth or gold standard

answers (e.g., correctness criteria), the prompt requests the model to also extract the gold label g_k from d for that criterion. This provides *explicit grounding* on the pre-training corpus. For criteria where there is no ground truth (e.g., creativity criteria), we prompt the model to utilize d when generating the criterion description, which serves as *implicit grounding* on the pre-training corpus.

In addition, the rubric must be **discriminative**: the model is prompted to generate only criteria that meaningfully distinguish high-quality responses from low-quality ones in the given response pool $\{y_{ref}\} \cup \{y_j\}_{j=1}^{J}$. Note that POP includes a d-privileged reference response y_{ref} in the rubric prompt. The rationale is that we want to discard examples where all candidate responses are highly similar and provide no meaningful contrastive signals. For these, y_{ref} serves as an extra reference point to avoid generating meaningless criteria that focus on local differences between candidate responses but do not correlate with global quality. Consequently, these examples receive the same score for all their candidate responses and will be filtered out in the filtering stage.

Response Grading (Grader). π_{ref} is asked via prompt P^{grader} to grade each candidate response y_j and produce an evaluation report e_j using the rubric r: $e_j \sim \pi_{ref}(\cdot|P^{grader}(x, r, y_j))$. e_j consists of individual evaluations e_j^k of y_j for each criterion r_k. For each e_j^k, we ask the model to give its thoughts on how well the model does according to r_k's description and the criterion-specific gold label g_k. Then, we ask it to produce a rating s_j^k of either 0 (Bad/Does not match gold), 1 (Medium/Partially matches gold), or 2 (Good/Fully matches gold), depending on the extent to which y_j satisfies r_k. If no valid score can be extracted for a criterion, we set its score to 0. Finally, we aggregate the scores into a final score $s_j \in [0.0, 2.0]$ as follows.

$$s_j = \frac{\sum_{k=1}^{K} w_k \cdot s_j^k}{\sum_{k=1}^{K} w_k} \tag{1}$$

3.2 Filtering, Pairing, and Training

Preliminaries. Direct Preference Optimization Rafailov et al. (2023) is an offline version of traditional policy gradient RL methods such as PPO Schulman et al. (2017). It avoids the need for an extra reward model, and trains the main policy to learn the reward landscape directly. Given a dataset $\{(x^i, y_w^i, y_l^i)\}_{i=1}^{N}$, where each example has a prompt x, a preferred response y_w, and a dispreferred response y_l, it initializes the main policy π_θ from the reference policy π_{ref} and trains it using the following objective.

$$L_{DPO}(\pi_\theta, \pi_{ref}) = -E_{(x,y_w,y_l)\sim S}\left[log\sigma\left(\beta log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \tag{2}$$

Filtering. After synthesizing the dataset, for each example, we filter out candidate responses that are malformed or have a malformed evaluation report. In addition, we filter out examples for which the query or rubric cannot be extracted, have an incorrect format, or have zero valid candidate responses.

Pairing. For each example that passes filtering, we pick the candidate response with the highest score as the preferred response $y_w \leftarrow argmax_{y_j}(s_j)$ and the one with the lowest score as the dispreferred response $y_l \leftarrow argmin_{y_j}(s_j)$. Prior work shows that DPO exhibits length biases Lu et al. (2024); Huang & Goyal (2025). Therefore, we filter out pairs where y_w and y_l differ by more than 100 words in length (See Appendix C for a detailed discussion). We also filter out pairs where y_w and y_l have the same score. This gives the DPO training set $\{(x^i, y_w^i, y_l^i)\}_{i=1}^{N}$.

Training. We initialize the main policy from π_{ref} and then train it using the DPO objective in equation (2) on the synthesized dataset. We denote the trained model as $\pi_{trained}$.

4 Experiment Setup

Tasks. We experiment with three tasks: Long-form Healthcare QA, Creative Writing, and Instruction Following. For **Long-form Healthcare QA**, we use HC4 Maslenkova et al. (2025) as our pre-training corpus, which consists of more than 9.7 million medical articles. We evaluate on HealthBench Arora et al. (2025), which assesses LLMs' ability to address queries from either patients or health professionals. Each example is accompanied by rubrics created by human experts. To reduce evaluation cost, we randomly sample 500 examples from it and denote it as HealthBench500. For

Task	Model	#Qry (I)	#Rsp (J)	#Criteria (K)	s mean	s std	s_{y_w}	s_{y_l}
Healthcare QA	Base	1381	10.69	4.48	1.19	0.47	1.78	0.45
	Inst	2667	24.45	2.71	1.01	0.29	1.51	0.54
Creative Writing	Base	1822	8.91	4.84	1.41	0.58	1.96	0.47
	Inst	2521	28.95	2.96	1.66	0.21	1.95	1.21
Instruction Following	Base	1665	10.53	4.47	1.20	0.50	1.80	0.43
	Inst	2502	24.68	3.04	1.35	0.30	1.79	0.80

Table 1: Statistics of the synthesized datasets, after filtering. Base: Qwen-2.5-7B; Inst: Qwen-2.5-7B-Instruct; #Qry: Number of queries; #Rsp: Number of valid candidate responses per query; #Crtieria: Number of rubric criteria per query; s mean: Average response score per query; s std: Answer score standard deviation per query; s_{y_w}: Average score of preferred response; s_{y_l}: Average score of dispreferred response. Score range is from 0.0 to 2.0.

Creative Writing, we use 2 million book abstracts scraped from the Internet Skelebor (2022) as the pre-training corpus. We evaluate on Creative Writing V3 Paech (2025), which assesses LLMs' creative writing skills via pre-defined rubrics. For **Instruction Following**, we use OpenWebText Gokaslan et al. (2019) as the pre-training corpus, which consists of 8 million documents in the general domains. For evaluation, we use IFEval Zhou et al. (2023), a verifiable benchmark that measures strict adherence to writing constraints (e.g., respond within 100 words), and ArenaHard Zheng et al. (2023), an LLM-as-a-judge benchmark measuring general instruction following abilities for diverse user queries (write emails, suggest weekend fun ideas, etc.). We use GPT-4.1-mini Team (2023) as the judge model for LLM-as-a-judge benchmarks.

OOD Evaluations. To ensure that our trained model's performance does not degrade on other tasks, especially on verifiable tasks, we evaluate POP on ten out-of-distribution (OOD) benchmarks: on scientific reasoning (GPQA-Diamond Rein et al. (2024)); math reasoning (GSM8K Cobbe et al. (2021); Math500 Hendrycks et al. (2021); AIME2024; AIME2025); and factoid QA (NaturalQueries Kwiatkowski et al. (2019); TriviaQA Joshi et al. (2017); TruthfulQA Lin et al. (2022); MMLU-Pro Wang et al. (2024)); MedMCQA Pal et al. (2022). We use 0-shot prompting for the benchmarks.

Models. We experiment with both a pretrained base model (Qwen-2.5-7B Team (2024)) and an instruction-finetuned model (Qwen-2.5-7B-Instruct Team (2024)) as our reference model π_{ref}. We also experiment with models from other families in Appendix G.

Sampling. We sample $I = 4096$ queries (except for Creative Writing, where we set $I = 8192$), each with $J = 32$ initial candidate responses, and ask the model to create at most 5 rubrics per query (i.e., $K = 5$). The statistics of the resulting synthesized datasets are shown in Table 1.

Training. We train the model for 1 epoch, with learning rate $1e^{-6}$, $\beta = 0.01$, and batch size 16. We use AdamW optimizer Loshchilov & Hutter (2019). See details in Appendix A.

Baselines. We compare $\pi_{trained}$ against (1) the reference model π_{ref}; (2) a model that is continuously pretrained on the same pre-training text used to ground our pipeline (**CPT on** D); (3) a model that is supervise-finetuned on the same queries, but with reference responses (**SFT on** y_{ref}). Prior rubric-based RL methods Gunjal et al. (2025); Viswanathan et al. (2025); Bi et al. (2025); Zhou et al. (2025a); He et al. (2025); Rezaei et al. (2025); Huang et al. (2025b) use supervision from large frontier models or humans to obtain input queries, rubrics, or scores, and are not comparable to POP.

5 Results and Analysis

We first discuss the main results in § 5.1, then evaluate the trained models on OOD benchmarks in § 5.2. We also analyze the synthesized training datasets in § 5.3 and conduct various ablation studies in § 5.4. Finally, we analyze the correctness of response ranking by our rubric evaluation in § 5.5.

5.1 Main Results

We show the main results in Table 2. Across different domains, we observe a consistent and statistically significant (See Appendix D) performance increase from POP over the reference Qwen models. In addition, both continuous pre-training on the grounding documents D (CPT on D) and supervised fine-tuning on the reference response y_{ref} (SFT on y_{ref}) underperform our methods

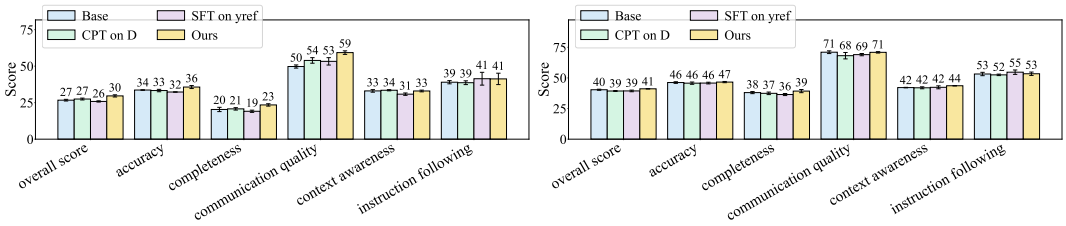


Figure 2: Per-aspect results on HealthBench500. Left: Qwen-2.5-7B; Right: Qwen-2.5-7B-Inst.

across all models and tasks, except for Qwen-2.5-7B on IFEval. In many cases, CPT and SFT even underperform the reference model. We suspect that this is due to a non-perfect train-test distribution match and further discuss it in Appendix E.

	Healthcare QA	Creative Writing	Instruction Following	
	HealthBench500	Creative Writing V3	IFEval	ArenaHard
Qwen-2.5-7B	26.71 ± 0.33	42.31 ± 0.47	31.79 ± 0.52	12.26 ± 0.19
CPT on D	27.49 ± 0.37	41.42 ± 0.52	33.64 ± 0.99	13.99 ± 0.16
SFT on y_{ref}	25.85 ± 0.24	41.93 ± 0.48	**37.46** ± 0.88	13.73 ± 0.45
POP	**29.64** ± 0.43	**44.31** ± 0.71	36.72 ± 0.70	**16.20** ± 0.17
Qwen-2.5-7B-Inst	40.35 ± 0.25	59.82 ± 0.37	70.30 ± 0.59	50.08 ± 0.82
CPT on D	39.31 ± 0.21	53.91 ± 0.26	69.87 ± 0.57	47.34 ± 0.53
SFT on y_{ref}	39.38 ± 0.38	58.25 ± 0.24	70.42 ± 0.15	45.87 ± 0.59
POP	**41.12** ± 0.08	**62.64** ± 0.91	**72.34** ± 0.50	**53.78** ± 0.23

Table 2: Main results. HealthBench and Creative Writing V3: we report standard rubric scores from original benchmarks; IFEval: We report the strict-prompt results, according to common practice Team (2024); Zhou et al. (2025a). ArenaHard: We report the win rate against GPT-4. We repeat all evaluations three times and report the standard errors. See Appendix D for 95% Confidence Intervals.

Healthcare QA. On the pretrained model (Qwen-2.5-7B), we observe a 3% increase on HealthBench500. The performance gains on the instruction-tuned model (Qwen-2.5-7B-Inst) are only 1%. We suspect this is due to misalignments between our synthesized training data and the evaluation data from HealthBench. In Figure 2, we show finer-grained results on benchmark-defined answer aspects. For Qwen-2.5-7B, performance increases notably on all axes except context awareness. For Qwen-2.5-7B-Inst, performance slightly increases in accuracy, completeness, and context awareness, but remains the same in communication quality and instruction following.

Creative Writing. On Creative Writing V3, POP gets a 2% increase for Qwen-2.5-7B and a 3% increase for Qwen-2.5-7B-Inst. Per-criteria results can be found in Appendix F.

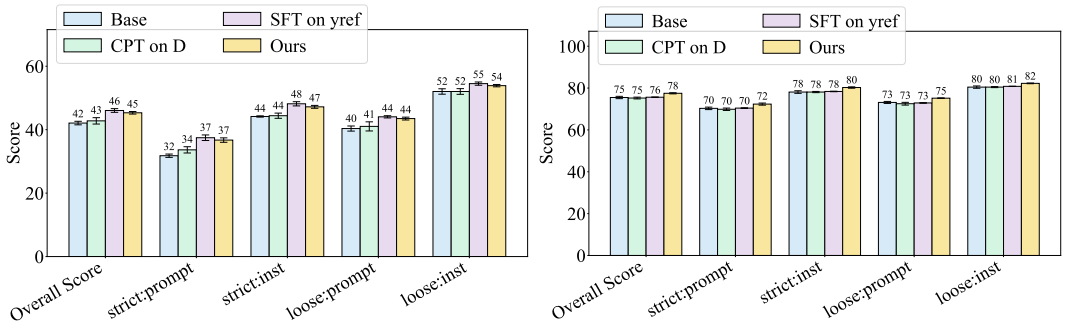


Figure 3: Per-axis results on IFEval. Left: Qwen-2.5-7B; Right: Qwen-2.5-7B-Inst.

Instruction Following. We first discuss results on IFEval. We observe a 5% increase for Qwen-2.5-7B and 2% increase for Qwen-2.5-7B-Inst. Detailed breakdown in Figure 3 indicates a consistent

	MedMCQA	NQ	TvQA	TfQA	MMLU-P	GPQA-D	GSM	Math	Aime24	Aime25
Qwen-2.5-7B										
Base	52.95	14.36	17.93	48.07	30.10	33.84	78.77	61.15	6.82	3.07
CPT on D	**54.00**	14.01	16.96	48.05	34.71	33.84	79.44	61.00	**7.76**	3.28
SFT on y_{ref}	53.81	**15.30**	16.55	48.62	43.78	31.82	78.05	59.60	6.09	3.33
POP	53.86	14.65	**18.96**	**49.81**	**44.75**	**35.35**	**79.98**	**64.00**	**7.76**	**4.32**
Qwen-2.5-7B-Inst										
Base	55.94	**15.12**	18.49	62.51	**56.57**	**36.36**	**90.64**	75.55	**12.40**	8.07
CPT on D	**56.99**	14.60	**22.80**	62.34	56.52	33.33	90.24	74.85	11.56	7.40
SFT on y_{ref}	56.78	14.83	17.54	**62.71**	56.20	33.84	90.16	**75.75**	11.25	7.55
POP	56.54	14.01	18.99	62.35	56.49	33.84	90.52	**75.75**	12.03	**8.23**

Table 3: OOD Results for models trained on the healthcare datasets. NQ: NaturalQueries; TvQA: TriviaQA; TfQA: TruthfulQA; MMLU-P: MMLU-Pro; GPQA-D: GPQA-Diamond; GSM: GSM8K; Math: MATH500. We use 0-shot evaluation.

performance increase under both strict and loose evaluations. On ArenaHard, POP improves the win rate against GPT-4 by 4% for Qwen-2.5-7B and 4% for Qwen-2.5-7B-Inst.

5.2 Results on OOD Benchmarks

We present the OOD results for models trained on Healthcare QA datasets in Table 3. OOD results for models trained on the Creative Writing and Instruction Following datasets can be found in Appendix H. Across models and tasks, we see no performance drop on OOD benchmarks.

We observe a slight increase across benchmarks for Qwen-2.5-7B. Notably, MMLU-Pro results increase from 30.10% to 44.75%. Our analysis in Appendix H shows that this is due to the model's better formatting and instruction-following abilities. Average results change from 34.71% to 37.34%. On Qwen-2.5-7B-Inst, results slightly drop on NaturalQueries and GPQA-Diamond, but remains largely the same on other benchmarks. Average results change from 43.17% to 42.88%. We also highlight that results on MedMCQA increase by around 1% for both models, suggesting that training on our long-form datasets slightly helps performance on short-form tasks in the same domain. On Qwen-2.5-7B, in terms of average OOD scores, we (37.34%) also slightly outperform CPT on D (35.30%) and SFT on y_{ref} (35.70%). On Qwen-2.5-7B-Inst, our results (42.88%) are similar to the baselines (43.06% for CPT on D and 42.66% for SFT on y_{ref}).

5.3 Statistics of the Synthesized Dataset

Below we analyze the types of queries, rubrics, and responses synthesized by POP. We use the Healthcare $\text{QA}_{\text{Qwen-2.5-7B}}$ dataset as an example. See details in Appendix I.

Queries. In Figure 10 (a), we show the common topics for the synthesized queries. Healthcare Management (21%) and Public Health (19%) are the two most common topics, which also have the strongest overlap with queries from HealthBench. Since a considerable portion of our pre-training corpus Maslenkova et al. (2025) is medical articles from PubMed, the rest of the queries tend to focus on professional medical knowledge and research.

Rubrics. We group the rubric criteria into meta-criteria and show the composition in Figure 10 (b). Relevance (25%), Accuracy (22%), and Clarity (18%) make up two-thirds of the rubric criteria. Other major categories include Depth (11%), Biological Understanding (8%), and Originality (6%). This suggests that our rubric values both objective and subjective criteria.

Responses and Scores. In Figure 4, we show the distribution of response scores, grouped by queries. According to the rubric and scores generated by the model itself, more than 50% of queries have their average scores above 1.0. However, as we will show in § 5.5, these scores are not perfect without the pairing step. Therefore, we additionally ask a stronger teacher model $\pi_{teacher}$ (GPT-4o-mini) to generate its own rubrics and scores on the same queries and responses (right side). The teacher distribution moves toward the lower end. Over 50% of queries have an average score below 1.0.

Nonetheless, we observe a large standard deviation of over 0.3 across query groups under both models. This suggests that POP is able to synthesize queries of appropriate difficulties, so that the model generates responses of varying quality, and can give meaningful contrastive signals.

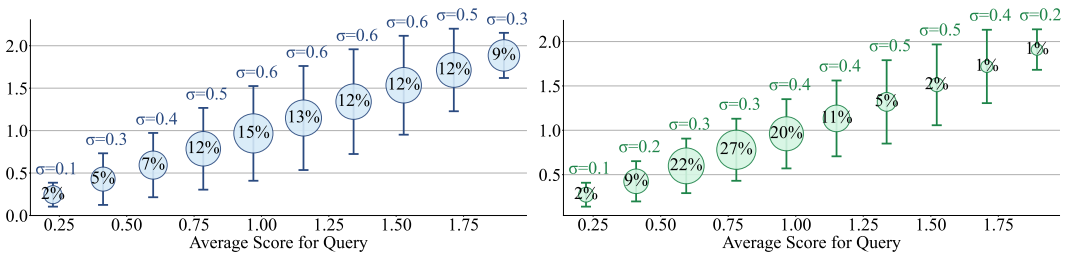


Figure 4: Response score distribution for healthcare$_{\text{Qwen-2.5-7B}}$ dataset. Left: scores given by the model itself; Right: scores given by a stronger teacher model; We group the candidate responses by queries and compute the average and standard deviation of response scores for each query. We then partition the queries into 10 bins according to average response scores. x-axis is the average score for the queries in that bin; error bar indicates standard deviations of scores for the queries in that bin, averaged by number of queries; bubble size indicates the percentage of queries that fall into that bin.

5.4 Ablation Study

We remove or replace various components of POP. This includes (1) revoking access to pre-training text d when generating the rubric (`Eval w/o` D); (2) revoking our entire pipeline's access to d altogether (`w/o` D); (3) replacing our pointwise rubric grader with a pairwise judge that compares two responses according to our rubric (`w/ pairwise judge`); (4) removing the access to rubrics for our judge, but still grounding it on d (`w/o rubric`). See Appendix J.1 for details.

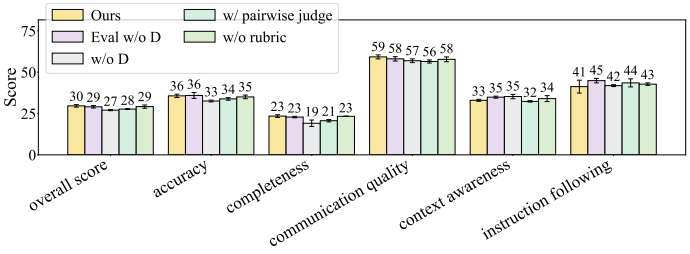


Figure 5: Ablation Results on HealthBench500.

Figure 5 shows the results for the ablation study on the Healthcare$_{\text{Qwen-2.5-7B}}$ dataset. Removing grounding of our rubric evaluation on the pre-training document reduces the overall performance to 29%. We suspect this to be a result of an insufficient generation-verification gap and an increase in reward hacking, which we will analyze in § 5.5. Removing our pipeline's access to D entirely (w/o D) further hurts performance, since in this case, queries are synthesized without grounding, and likely degrade in terms of quality and diversity. Grading responses without rubrics slightly hurts the performance, and we argue that rubrics make evaluation more interpretable. Surprisingly, using a pairwise judge is not better than ours, which we will analyze in Appendix J.2.

5.5 Response Ranking Analysis

Without external input, using the same model to verify its own responses risks an insufficient generation-verification gap Song et al. (2025). In our case, the rubric generated by our model may not distinguish high-quality responses from low-quality ones. As a result, the ranking of responses according to the rubric will be inconsistent with the "true quality" ranking. Choosing responses based on this ranking for training encourages undesired behavior, a phenomenon known as **reward hacking**. In our case, since we only choose the highest-ranked y_w and lowest-ranked y_l responses for DPO training, reward hacking happens when y_w is worse than y_l in terms of "true qualities".

To measure "true qualities", we use a stronger teacher model $\pi_{teacher}$ as a proxy. In particular, we use our prompts to ask $\pi_{teacher}$ to generate a new set of rubrics and scores for the same queries and responses from π_{ref}. We define the rankings using this stronger model as the *gold rankings*. We want to know (1) the correlation between POP's rankings and gold rankings, and (2) whether grounding our rubric on pre-training document d helps. For (2), we revoke access to D and regenerate the rubrics and scores with π_{ref} (`Eval w/o` D). We compare the new rankings with our original rankings (`Eval w/` D). Both use the gold rankings as a reference. We use percentile ranking, so a rank of x means the response is ranked at the top x% among the responses to the same query.

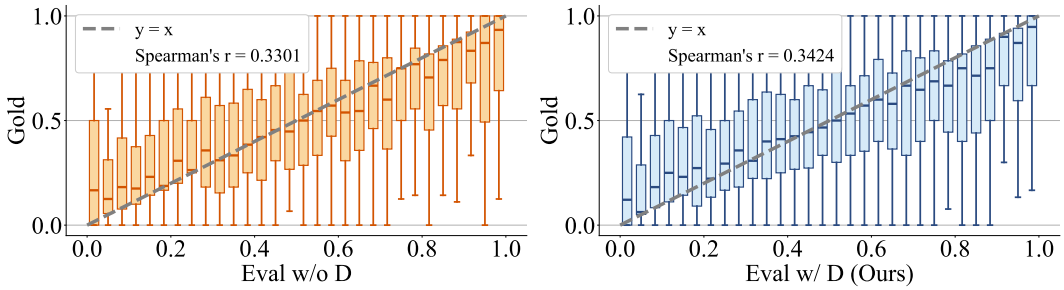


Figure 6: Correlation between our rankings and gold rankings on the $\text{Healthcare}_{\text{Qwen-2.5-7B}}$ dataset. x-axis: Ranking of responses from our model, with (Right) or without (Left) access to D. y-axis: For responses that are ranked at the top x% among the responses to the same query according to our model, the distribution of their gold rankings. Spearman's r: Spearman's ranking correlation.

Full correlation is imperfect. Figure 6 shows the correlation between our rankings and gold rankings. Evaluation without grounding (`Eval w/o` D) achieves a moderate positive correlation of 0.3301 with gold rankings. Grounding on pre-training text (`Eval w/` D) gives a slightly stronger correlation of 0.3424, but still far from perfect.

$y_w - y_l$ correlation is strong, leading to correct training signals. Since our DPO algorithm only uses the highest and lowest ranked responses y_w and y_l for training, a more relevant metric for POP is the pairwise ranking accuracy. That is, the percentage of cases where y_w is truly better than y_l according to gold rankings. We show this in Table 4.

	$\%(rk^{gold}_{y_w} \leq rk^{gold}_{y_l})$
Eval w/o D	82.04
Eval w/ D (Ours)	85.14

Table 4: Pairwise Ranking accuracy.

Both settings achieve accuracies above 80%, and `Eval w/` D achieves an accuracy of 85%. This suggests (1) even though global correlation is imperfect, taking the extremes according to our own rankings still gives mostly correct training signals; (2) Grounding rubric evaluation on the pre-training corpus gives stronger correlation with gold rankings and helps reduce reward hacking.

In Table 5, we further list the detailed statistics of y_w and y_l under both `Eval w/o` D and `Eval w/` D. For both settings, we compute the average rankings and scores of y_w and y_l under our model and the teacher model. In both settings, y_w is ranked and scored significantly higher than y_l by the teacher model. In addition, the rank difference between y_w and y_l is larger in the `Eval w/` D setting, showing the benefits of grounding on pre-training text. We provide further analysis and methodology in Appendix K and conduct a human study in Appendix L.

	Verifier	rk_{y_w}	rk_{y_l}	Δrk	s_{y_w}	s_{y_l}	Δs
Eval w/o D	π_{ref}	0	100	100	1.78	0.47	1.31
	$\pi_{teacher}$	26.59	75.35	48.76	1.01	0.63	0.38
Eval w/ D (Ours)	π_{ref}	0	100	100	1.79	0.51	1.28
	$\pi_{teacher}$	23.59	76.50	52.91	1.04	0.63	0.41

Table 5: Rankings and scores of y_w and y_l from Eval w/o D and Eval w/ D, according to either π_{ref} or stronger model.

6 Conclusion

We introduce POP, a rubric-based self-play post-training framework on open-ended tasks. POP synthesizes query-response-rubric triples and grades each response using the rubrics. To mitigate reward hacking, we give privileged access to pretraining text during rubric generation and only take the responses with the highest and lowest scores for DPO training. Experiment results suggest that POP bootstraps effective training signals across different tasks. The design of POP is meant to be generalizable, and we hope it highlights a direction toward cheap and effective post-training for LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, as part of our proposed method. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, or interpreting results. Helping to develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs and proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images. We have reviewed all AI-assisted work. The synthesized datasets are manually inspected by the authors in terms of their relevance to the target task and the figures are manually examined in terms of their faithfulness to the experiments results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code for sampling, filtering, pairing, and training in the attached supplementary materials. The supplementary materials also contains dataset examples (before pairing). For all experiments, the final hyperparameters as well as their tuning efforts are described in Appendix A. The computation resources used are described in Appendix P.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Baolong Bi, Shenghua Liu, Yiwei Wang, Siqian Tong, Lingrui Mei, Yuyao Ge, Yilong Xu, Jiafeng Guo, and Xueqi Cheng. Reward and guidance through rubrics: Promoting exploration to improve multi-domain reasoning. *arXiv preprint arXiv:2511.12344*, 2025.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K. Wong. Spc: Evolving self-play critic via adversarial games for llm reasoning. *arXiv preprint arXiv:2504.19162*, 2025.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Zheng Yuan, Yong Dai, Lei Han, Nan Du, and Xiaolong Li. Self-playing adversarial language game enhances llm reasoning. *arXiv preprint arXiv:2404.10642*, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

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. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint:arXiv:2512.02556*, 2025.

Jacob Dineen, Aswin Rrv, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, and Ben Zhou. Qa-lign: Aligning llms through constitutionally decomposed qa. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

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.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G. Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Awadalla, and Manaal Faruqui. Advancedif: Rubric-based benchmarking and reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*, 2025.

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 *Advances in Neural Information Processing Systems*, 2021.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025a.

Chengyu Huang and Tanya Goyal. Dcrm: A heuristic to measure response pair quality in preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*, 2025b.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, art. arXiv:1705.03551, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics, Volume 7*, 2019.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations*, 2019.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Thang Luong and Edward Lockhart. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad, 2025.

Svetlana Maslenkova, Clement Christophe, Marco AF Pimentel, Tathagata Raha, Muhammad Umar Salman, Ahmed Al Mahrooqi, Avani Gupta, Shadab Khan, Ronnie Rajan, and Praveenkumar Kanithi. Building trust in clinical llms: Bias analysis and dataset transparency. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Samuel J Paech. Eq-bench creative writing benchmark v3. https://github.com/EQ-bench/creative-writing-bench, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. *arXiv preprint arXiv:2503.23037*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *arXiv preprint arXiv:2305.18290*, 2023.

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

Yi Ren and Danica J. Sutherland. Learning dynamics of llm finetuning. In *Proceedings of International Conference on Learning Representations*, 2025.

Databricks AI Research. Karl: Knowledge agentsvial reinforcement learning. *arXiv preprint:arXiv:2603.05218*, 2026.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025.

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

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

Harman Singh, Xiuyu Li, Kusha Sareen, Monishwaran Maheswaran, Sijun Tan, Xiaoxia Wu, Junxiong Wang, Alpay Ariyak, Qingyang Wu, Samir Khaki, Rishabh Tiwari, Long Lian, Yucheng Lu, Boyi Li, Alane Suhr, Ben Athiwaratkun, and Kurt Keutzer. V1: Unifying generation and self-verification for parallel reasoners. *arXiv preprint arXiv:2603.04304*, 2026.

Skelebor. Book titles and abstracts. https://huggingface.co/datasets/Skelebor/book_titles_and_descriptions, 2022.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *Proceedings of International Conference on Learning Representations*, 2025.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, and Will Dabney Yong Cheng. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

OpenAI Team. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zhengxin Zhang, Chengyu Huang, Aochong Oliver Li, and Claire Cardie. Better llm reasoning via dual-play. *arXiv preprint arXiv:2511.11881*, 2025.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

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. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025a.

Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. Self-challenging language model agents. *arXiv preprint arXiv:2506.01716*, 2025b.

A HYPERPARAMETER

We show the list of hyperparameters in Table 6.

Stage	Parameter	Value
Sampling	Max Context Length	32768
	$\|d\|$	$\in [50, 1024]$
	Proposer Max New Tokens	6144
	Proposer Temperature	1.0
	Proposer Top P	1.0
	Solver Max New Tokens	6144
	Solver Temperature	1.0
	Solver Top P	1.0
	Rubric Gen Max New Tokens	8192
	Rubric Gen Temperature	0.0
	Rubric Gen Top P	1.0
	Ans Grading Max New Tokens	4096
	Ans Grading Gen Temperature	0.0
	Ans Grading Gen Top P	1.0
	# Ques (I)	4096
	# Ques (I) Creative Writing	8192
	# Ques per d	1
	# Ans per Ques (J)	32
	Max # Rub Criteria per Ques (K)	5
Continuous Pre-training (CPT on D)	Scheduler	Cosine
	Warmup Ratio	0.1
	Optimizer	AdamW
	Learning Rate	5e-6
	Number of Epochs	1
	Batch Size	64
Supervised-Finetuning (SFT on y_{ref})	Scheduler	Cosine
	Warmup Ratio	0.1
	Optimizer	AdamW
	Learning Rate	1e-6
	Number of Epochs	1
	Batch Size	64
DPO (POP)	Scheduler	Cosine
	Warmup Ratio	0.1
	Optimizer	AdamW
	Learning Rate	1e-6
	Beta	0.01
	Number of Epochs	1
	Batch Size	16

Table 6: Hyperparameters

For Continuous Pre-training and Supervise-Finetuning, we initially use a learning rate of 2e-5 but observe significant performance degradation, so we search for the optimal learning rate among $\{1e^{-6}, 2e^{-6}, 5e^{-6}, 1e^{-5}, 2e^{-5}\}$, using a held-out subset of 100 examples from HealthBench as the validation set. For our DPO approach, we search the learning rate among $\{1e^{-6}, 3e^{-6}, 5e^{-6}\}$ and beta among $\{0.1, 0.01\}$.

For rubric generation, to prevent context length overflow, we select ten candidate responses to put into the rubric generation prompt. In particular, we sort the responses by length and then pick one from every $\frac{\#Ans}{10}$ so that the rubric generator sees responses of varying length.

A.1 EVALUATION

Main Evaluation. We use the default hyperparameters for the evaluations whenever such defaults are provided. For HealthBench, when we randomly sample 500 examples from the original bench-

marks, we keep the proportion of hard examples fixed at 0.2. We use a sampling temperature of 0.05 and a max output length of 4000 for our model, and a temperature of 0.1 and a max output length of 4000 for the judge. For Creative Writing V3, we use a sampling temperature of 0.7 and a max output length of 4000 for our model and a temperature of 0.1 and a max output length of 4096 for the judge. For IFEval, we use a sampling temperature of 0.7 and a max context length of 4096. For AreaHard, we use a sampling temperature of 0 and a max output length of 2048 for our model and a temperature of 0 and a max output length of 16000 for the judge.

OOD Evaluation. We use lighteval Habib et al. (2023) for OOD Evaluations. Evaluations are 0-shot across different benchmarks. For math benchmarks (GSM8K, Math500, AIME2024, AIME2025), we use a sampling temperature of 0.6, a max context length of 32768, and top p = 0.95. We report avg@4 for GSM8K and Math500 and avg@64 for AIME2024 and AIME2025. For other benchmarks, we use a sampling temperature of 0 and a max context length of 4096.

Analysis. For our analysis with a stronger teacher model $\pi_{teacher}$, we set $\pi_{teacher}$ to GPT-4o-mini.

B POP ALGORITHM

We show the pseudo code of POP in Algorithm 1.

C DISCUSSION ON LENGTH BIASES

The length constraint ($||y_w| - |y_l|| \leq 100$ words) is introduced primarily to reduce the length bias of the DPO algorithms, not the length bias of the rubrics. In our training dataset, Healthcare QA with Qwen2.5-base, before the length-difference filtering, we observe slightly more cases where $|y_w| > |y_l|$ (1750/3171=55.19%) than cases where $|y_w| < |y_l|$ (1414/3171=44.59%). This ratio is not that different after the length filtering (51.56%/47.94%). Before the length filtering, the average length of $|y_w|$ (318.70) and $|y_l|$ (295.23) are also similar to each other, which are well within the constraint of 100 words. This shows that there is no systematic length bias in the training data both before and after applying the length constraints. Therefore, our rubric does not suffer from severe length bias. What changes after length filtering is the elimination of cases where $|y_w|$ is significantly longer than $|y_l|$. These cases induce the length bias of DPO, which is a well-known phenomenon Lu et al. (2024), where the model tends to produce longer responses after training. To test this, we run an ablation removing the length constraint. We observed a drop in downstream evaluation performance. This confirms that the length constraint is a necessary safeguard for preference learning, not a patch for rubric unreliability.

Algorithm 1 POP Pipeline

Require: Reference LLM π_{ref}; Pre-training Corpus D; #Queries I; #Responses /Q J; #Max Rubric Citeria/Qus K; Task t; Query Synthesis Prompt $P_t^{proposer}$; Response Generation Prompt P_t^{solver}; Rubric Generation Prompt P^{rubric}; Response Grading Prompt P^{grader}

// Sampling
Initialize RawDataset $\leftarrow \{\}$
for $i = 1$ to I **do**
    // Query Synthesis
    Sample $d \sim D$
    Sample $(x, y_{ref}) \sim \pi_{ref}(\cdot|P_t^{qus}(d))$
    // Response Generation
    Sample $\{y_j\}_{j=1}^{J} \sim \pi_{ref}(\cdot|P_t^{ans}(x))$
    // Rubric Generation
    Draw sample responses $Y_{rubric} \subset (\{y_j\}_{j=1}^{J})$ ▷ Prevent context overflow
    Sample $r \sim \pi_{ref}(\cdot|P^{rub}(d, x, y_{ref}, Y_{rubric}))$ with $|r| \leq K$
    // Response Grading
    for $j = 1$ to J **do**
        Sample $e_j \sim \pi_{ref}(\cdot|P^{grade}(x, r, y_j))$
        $\{w_k\}_{k=1}^{K'} \leftarrow$ GetWeights(r)
        $\{s_j^k\}_{k=1}^{K'} \leftarrow$ GetScores(e)
        $s_j \leftarrow \frac{\sum_{k=1}^{K'} w_k \cdot s_j^k}{\sum_{k=1}^{K} w_k}$
    end for
    RawDataset $\leftarrow$ RawDataset $\cup \{(d, x, y_{ref}, \{y_j\}_{j=1}^{J}, r, \{(e_j, s_j)\}_{j=1}^{J})\}$
end for
// Filtering
FilteredDataset $\leftarrow \{\}$
for example in RawDataset **do**
    validIndices $\leftarrow \{\}$
    for j in $\{1, \cdots, J\}$ **do**
        if ResponseValid(y_j) and EvaluationValid(e_j) **then** ▷ Not extractable, format incorrect, etc.
            validIndices $\leftarrow$ validIndices $\cup \{j\}$
        end if
    end for
    ResetResponsesAndEvaluations(example, $\{y_j\}_{j \in \text{validIndices}}$, $\{e_j\}_{j \in \text{validIndices}}$)
    isvalid $\leftarrow True$
    for component in examples **do**
        if not CheckValidity(component) **then** ▷ Not extractable, format incorrect, etc.
            isvalid $\leftarrow False$
            Break
        end if
    end for
    if isvalid **then**
        FilteredDataset $\leftarrow$ FilteredDataset $\cup \{\text{example}\}$
    end if
end for
// Pairing
DPODataset $\leftarrow \{\}$
for example in FilteredDataset **do**
    $x \leftarrow$ GetQuery(example)
    $\{(y_j, s_j)\}_{j=1}^{J'} \leftarrow$ GetResponsesAndScores(example)
    $y_w \leftarrow argmax_{y_j}(s_j)$
    $y_l \leftarrow argmin_{y_j}(s_j)$
    if $y_w > y_l$ and $||y_w| - |y_l|| \leq 100$ **then**
        DPODataset $\leftarrow$ DPODataset $\cup \{(x, y_w, y_l)\}$
    end if
end for
// Training
$\pi_{trained} \leftarrow DPO(\pi_{ref}, \text{DPODataset})$
return $\pi_{trained}$

Benchmark	Model	π_{ref}	π_{ref} 95% CI	$\pi_{trained}$	$\pi_{trained}$ 95% CI	Stat. Significant
HealthBench	Base	26.71 $_{\pm 0.33}$	[26.07, 27.36]	29.64 $_{\pm 0.43}$	[28.79, 30.49]	Yes (No overlap)
	Inst	40.35 $_{\pm 0.25}$	[39.86, 40.84]	41.12 $_{\pm 0.08}$	[40.96, 41.28]	Yes (No overlap)
CreativeWritingV3	Base	42.31 $_{\pm 0.47}$	[41.39, 43.23]	44.31 $_{\pm 0.71}$	[42.92, 45.70]	Yes (Slight overlap)
	Inst	59.82 $_{\pm 0.37}$	[59.08, 60.56]	62.64 $_{\pm 0.91}$	[60.86, 64.42]	Yes (No overlap)
IFEval	Base	31.79 $_{\pm 0.52}$	[30.76, 32.82]	36.72 $_{\pm 0.70}$	[35.35, 38.09]	Yes (No overlap)
	Inst	70.30 $_{\pm 0.59}$	[69.15, 71.45]	72.34 $_{\pm 0.50}$	[71.36, 73.33]	Yes (No overlap)
ArenaHard	Base	12.26 $_{\pm 0.19}$	[11.89, 12.63]	16.20 $_{\pm 0.17}$	[15.87, 16.53]	Yes (No overlap)
	Inst	50.08 $_{\pm 0.82}$	[48.46, 51.70]	53.78 $_{\pm 0.23}$	[53.34, 54.22]	Yes (No overlap)

Table 7: Statistical significance testing of the reference model π_{ref} and model trained with POP $\pi_{trained}$.

D Statistical Significance Test

We report the 95% confidence intervals of π_{ref} and $\pi_{trained}$ across three independent evaluation runs in Table 7. The improvements of POP are highly statistically significant. There is only a slight overlap in confidence intervals for Creative Writing with the base model and no overlap for the rest.

E Discussion on Train-test Distribution Match

We perform extensive hyperparameter tuning (searching learning rates $\{1e^{-6}, 2e^{-6}, 5e^{-6}, 1e^{-5}, 2e^{-5}\}$ as detailed in Appendix A) and find that both continuous pre-training on the grounding documents D (CPT on D) and supervised fine-tuning on the reference response y_{ref} (SFT on y_{ref}) consistently struggle. We suspect that the performance degradation is due to **non-perfect train-test distribution match**. Fundamentally, without prior knowledge of the testing benchmarks, it is impossible to synthesize training questions that are from the exact same distribution as the testing questions, regardless of specific training algorithms. In light of this, POP is intentionally designed to be generalizable, and so it has **minimal priors specific to any of the testing benchmarks**:

- **No Tailored Prompts:** Unlike many existing post-training or alignment methods, we do not design long, custom proposer, solver, or verifier prompts tailored to the specific evaluation benchmarks (e.g., HealthBench, Creative Writing V3, IFEval, ArenaHard). Instead, we use generic, task-level prompts for the proposer and solver (e.g., general healthcare QA, general creative writing, general instruction following), and global prompts for the verifier.
- **No Overlapping Data Distributions:** The pre-training corpora used for self-play (HC4, book abstracts, OpenWebText) are entirely separate from the evaluation benchmarks. POP does not require any training data from the exact same distribution as the test set.

On Healthcare QA, the resulting training distribution contains many in-depth medical research questions ($\geq$57% of the questions for base and $\geq$56% of the questions for inst) that are not tested in HealthBench. The non-perfect match between our training and testing distributions makes CPT/SFT less effective, especially with a relatively small training set. In contrast, DPO is more effective in this setting because the contrastive signals between high- and low-quality responses (clearer; more detailed; etc.) are often more generalizable and transferable to slightly out-of-distribution testing data.

This highlights why CPT on D/SFT on y_{ref} underperforms the base model and DPO. Nonetheless, we believe that SFT on y_{ref} could be highly performant if the training set size scales up significantly, but that is beyond our compute resources.

F Results on Creative Writing V3

We show the per-criterion results for Qwen-2.5-7B in Figure 7 and results for Qwen-2.5-7B-Inst in Figure 8. POP achieves consistent improvement over the reference models for both cases. Exceptions include "Unsurprising or Uncreative" and "Amateurish" for both models, and "Meandering", "Weak Dialogue", "Tell-Don't-Show", "Purple Prose", "Overwrought", "Incongruent Ending Positivity",

and "Unearned Transformations" for Qwen-2.5-7B-Inst. Continuous pre-training on D significantly degrades performance on Qwen-2.5-7B-Inst.

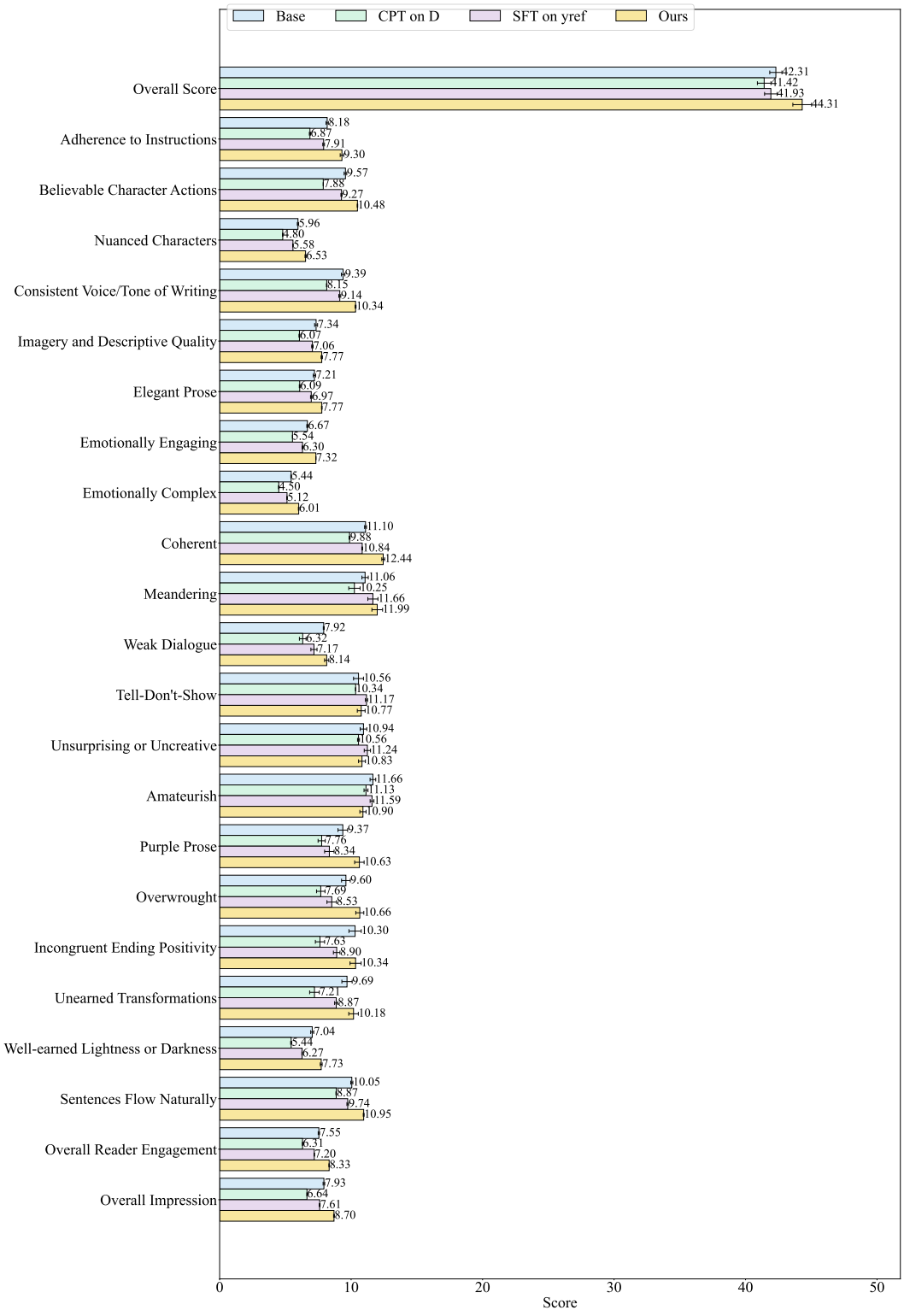


Figure 7: Qwen-2.5-7B Results on Creative Writing V3.

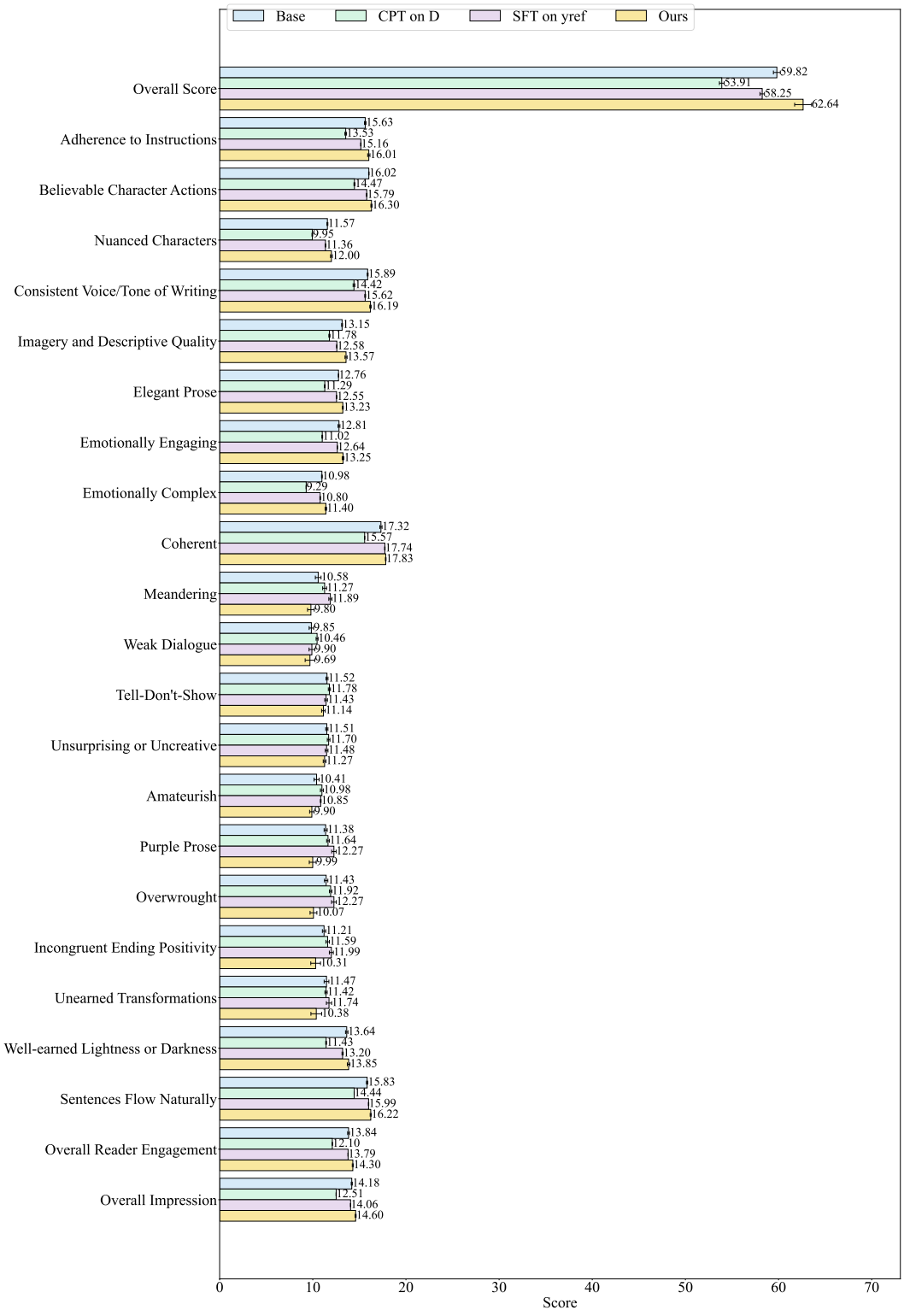


Figure 8: Qwen-2.5-7B-Inst Results on Creative Writing V3.

G RESULTS ON OTHER MODELS

We additionally experiment with a model from other families, **LLaMA3.1-8B-Instruct**, in the Healthcare QA setting. We use the same pretraining corpus and sample 4,096 queries and corresponding responses. After filtering, the final training dataset for DPO has 1,474 response pairs. We show the results on the HealthBench in Figure 9. On average, POP improves the performance from $36.02_{\pm 0.42}$ (95% CI = [35.19, 36.85]) to $38.36_{\pm 0.09}$ (95% CI = [38.19, 38.53]).

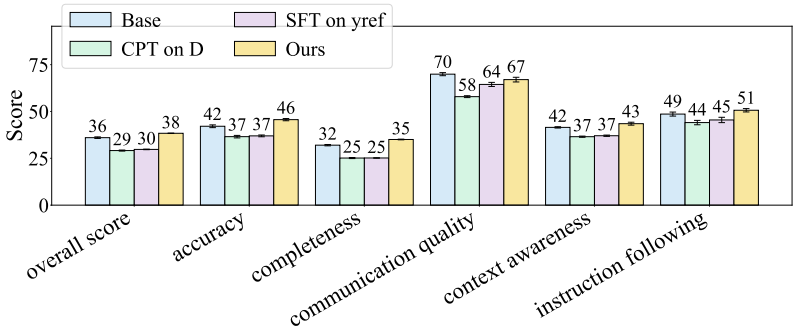


Figure 9: LLaMA3.1-8B-Instruct Results on HealthBench.

H RESULTS ON OOD BENCHMARKS

We show the OOD results of models trained on the Healthcare QA datasets in Table 8, Creative Writing datasets in Table 10, and Instruction Following datasets in Table 11. Both POP and the baselines are able to maintain the performance across different benchmarks compared with the reference model.

	MedMCQA	NQ	TvQA	TfQA	MMLU-P	GPQA-D	GSM	Math	Aime24	Aime25
Qwen-2.5-7B										
Base	52.95	14.36	17.93	48.07	30.10	33.84	78.77	61.15	6.82	3.07
CPT on D	**54.00**	14.01	16.96	48.05	34.71	33.84	79.44	61.00	**7.76**	3.28
SFT on y_{ref}	53.81	**15.30**	16.55	48.62	43.78	31.82	78.05	59.60	6.09	3.33
POP	53.86	14.65	**18.96**	**49.81**	**44.75**	**35.35**	**79.98**	**64.00**	**7.76**	**4.32**
Qwen-2.5-7B-Inst										
Base	55.94	**15.12**	18.49	62.51	**56.57**	**36.36**	**90.64**	75.55	**12.40**	8.07
CPT on D	**56.99**	14.60	**22.80**	62.34	56.52	33.33	90.24	74.85	11.56	7.40
SFT on y_{ref}	56.78	14.83	17.54	**62.71**	56.20	33.84	90.16	**75.75**	11.25	7.55
POP	56.54	14.01	18.99	62.35	56.49	33.84	90.52	**75.75**	12.03	**8.23**

Table 8: OOD Results for models trained on the models trained on the healthcare datasets. NQ: NaturalQuestions; TvQA: TriviaQA; TfQA: TruthfulQA; MMLU-P: MMLU-Pro; GPQA-D: GPQA-Diamond; GSM: GSM8K; Math: MATH500. We use 0-shot evaluation.

Notably, for the base model trained on the Healthcare QA task (Table 8), the results on MMLU-Pro increase from 30.10% to 44.75%. We analyze the behavioral differences between Qwen2.5-7B-Base and our Healthcare QA fine-tuned version on MMLU-Pro. Our model has far fewer runaway long generations. The base model tends to keep generating either new questions or answers after the initial answer. As shown in Table 9, almost all the cases where the base model answers wrongly but our model answers correctly are due to the aforementioned runaway generation behaviors. Therefore, the Pass@1 difference is due to **better answer formatting and**

Error Type	Contributions to Pass@1 Difference
Runaway generations that contain a correct answer but extracted prediction is wrong	9.21
Runaway generations that contain no correct answer	7.49
Other errors	-2.05

Table 9: Pairwise Ranking Accuracy.

instruction-following ability after finetuning rather than a sudden surge in core knowledge. We also observe a similar phenomenon on the base model in the Instruction Following setting (Table 11), where its performance on MMLU-Pro increases from 30.10% to 43.82%. The causes are the same: runaway responses, regardless of whether they contain the correct answers (9.57) or not (6.92), diminish after training.

	MedMCQA	NQ	TvQA	TfQA	MMLU-P	GPQA-D	GSM	Math	Aime24	Aime25
Qwen-2.5-7B										
Base	52.95	**14.36**	17.93	48.07	30.10	33.84	78.77	61.15	6.82	3.07
CPT on D	52.31	12.31	16.18	46.95	**36.90**	**35.35**	80.02	60.55	5.99	3.39
SFT on y_{ref}	**53.86**	14.30	**18.89**	**49.18**	30.33	31.82	**81.29**	**64.05**	7.76	3.54
POP	52.95	13.48	18.14	48.49	36.86	30.81	80.17	62.26	**7.86**	**4.32**
Qwen-2.5-7B-Inst										
Base	55.94	**15.12**	18.49	62.51	56.57	36.36	90.64	**75.55**	**12.40**	**8.07**
CPT on D	56.16	14.54	**19.35**	61.51	56.26	38.28	90.73	74.15	12.08	7.19
SFT on y_{ref}	**56.28**	15.01	18.35	**62.87**	56.49	**39.39**	**90.86**	75.15	11.35	6.46
POP	56.08	14.65	18.21	62.66	**56.62**	34.85	90.18	74.75	11.98	7.81

Table 10: OOD Results for models trained on the Creative Writing datasets.

	MedMCQA	NQ	TvQA	TfQA	MMLU-P	GPQA-D	GSM	Math	Aime24	Aime25
Qwen-2.5-7B										
Base	52.95	14.36	17.93	48.07	30.10	**33.84**	78.77	61.15	6.82	3.07
CPT on D	53.22	13.77	17.15	47.83	39.89	31.31	81.88	61.90	7.50	3.18
SFT on y_{ref}	**53.55**	**14.95**	16.48	**48.92**	**46.00**	**33.84**	81.07	60.80	5.68	2.71
POP	52.98	14.07	**18.86**	48.18	43.82	27.78	**83.26**	**64.25**	**7.92**	**4.69**
Qwen-2.5-7B-Inst										
Base	55.94	**15.12**	18.49	62.51	**56.57**	36.36	**90.64**	75.55	**12.40**	8.07
CPT on D	55.96	14.19	**22.16**	61.46	56.52	37.37	90.33	75.55	11.98	7.50
SFT on y_{ref}	56.08	15.12	16.60	**62.75**	56.14	31.31	90.45	75.40	11.25	6.51
POP	**56.47**	14.13	18.30	62.12	56.44	**38.38**	90.24	**76.35**	11.82	**9.84**

Table 11: OOD Results for models trained on the Instruction Following datasets.

I STATISTICS OF THE SYNTHESIZED DATASETS

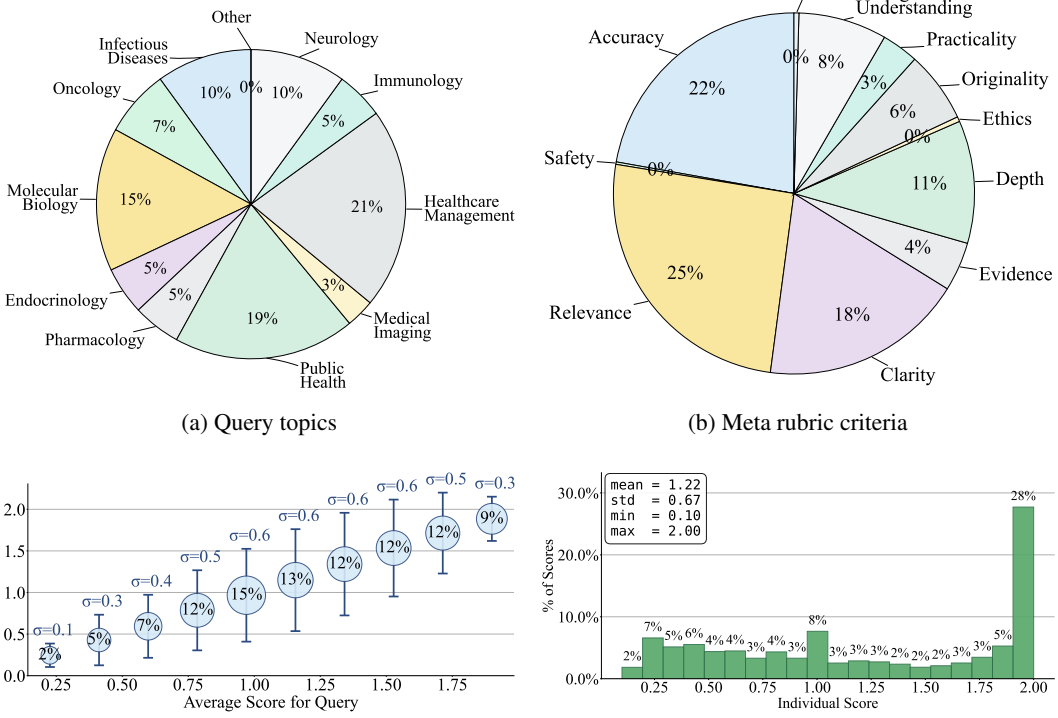
I.1 QUERIES AND RUBRICS CLASSIFICATION

Queries. For each task, we sample 100 examples from both Qwen-2.5-7B and Qwen-2.5-7B-Instruct. The queries for the resulting 200 examples are fed into GPT-4.1-mini to identify the common topics. We then ask GPT-4.1-mini again to categorize each query to exactly 1 of the topics.

Rubrics. We take the rubric criteria from the 200 examples used in the previous query topic analysis and feed them into GPT-4.1-mini to identify the meta-criteria. We then ask GPT-4.1-mini again to categorize each rubric criterion to exactly 1 of the meta-criteria.

I.2 HEALTHCARE QA

We show the statistics for our synthesized datasets $\text{HealthcareQA}_{\text{Qwen-2.5-7B}}$ in Figure 10 and $\text{HealthcareQA}_{\text{Qwen-2.5-7B-Inst}}$ in Figure 11.



(a) Query topics

(b) Meta rubric criteria

(c) Response score distribution, grouped by queries.

(d) Individual response score distribution.

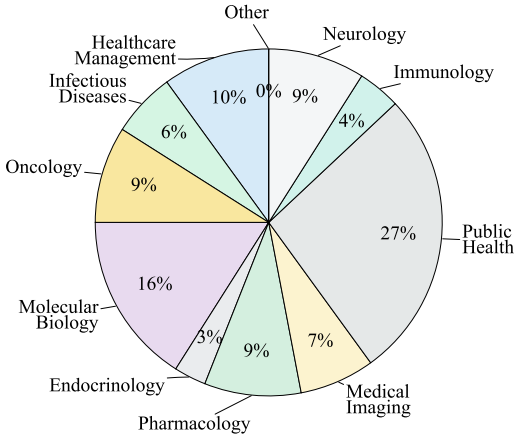
Figure 10: Statistics for Healthcare $\text{QA}_{\text{Qwen-2.5-7B}}$ dataset.

I.3 CREATIVE WRITING

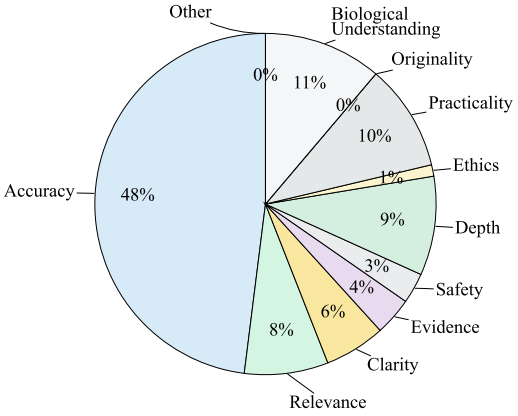
We show the statistics for our synthesized datasets $\text{CreativeWriting}_{\text{Qwen-2.5-7B}}$ in Figure 12 and $\text{CreativeWriting}_{\text{Qwen-2.5-7B-Inst}}$ in Figure 13.

I.4 INSTRUCTION FOLLOWING

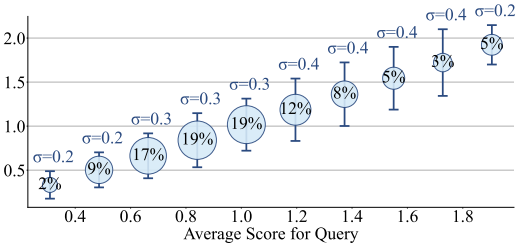
We show the statistics for our synthesized datasets $\text{InstructionFollowing}_{\text{Qwen-2.5-7B}}$ in Figure 14 and $\text{InstructionFollowing}_{\text{Qwen-2.5-7B-Inst}}$ in Figure 15. We highlight that we use the most general prompt (Figure 54) and pre-training corpus (OpenWebText) for this task, so the topic coverage is much



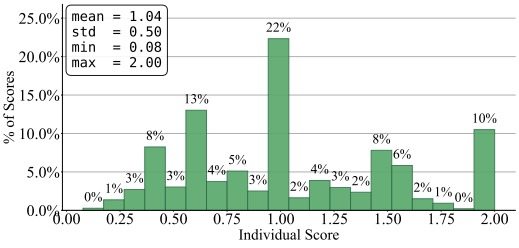
(a) Query topics



(b) Meta rubric criteria

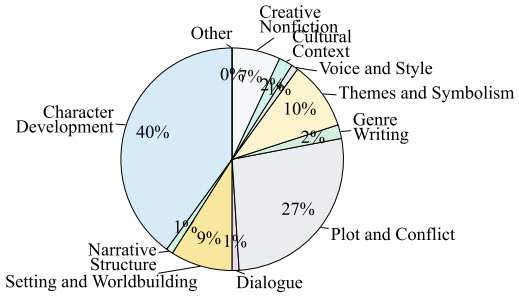


(c) Response score distribution, grouped by queries.

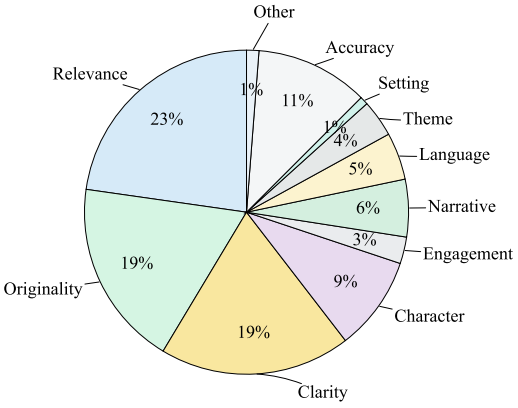


(d) Individual response score distribution.

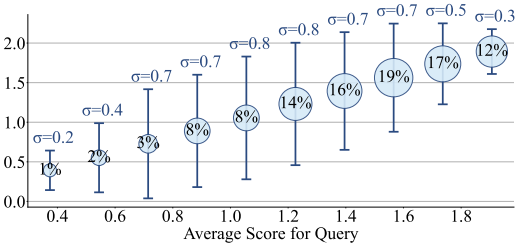
Figure 11: Statistics for Healthcare $QA_{\text{Qwen-2.5-7B-Inst}}$ dataset



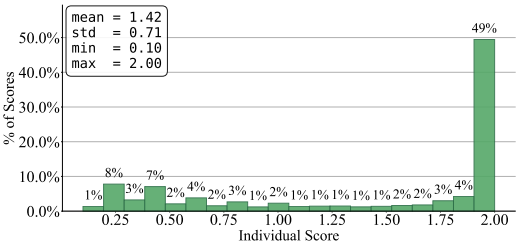
(a) Query topics



(b) Meta rubric criteria

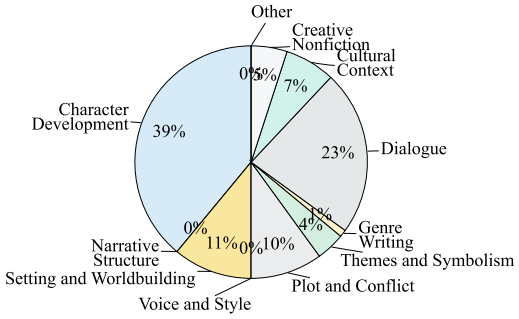


(c) Response score distribution, grouped by queries.

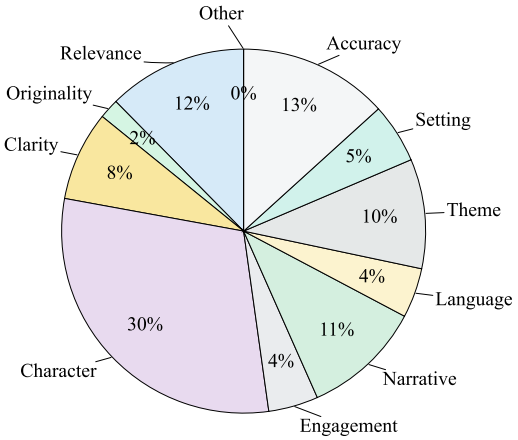


(d) Individual response score distribution.

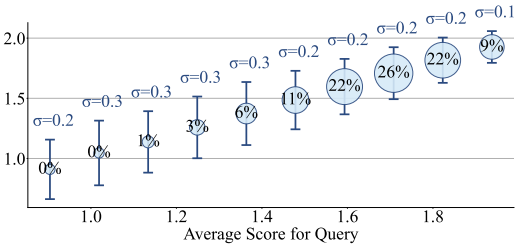
Figure 12: Statistics for Creative $\text{Writing}_{\text{Qwen-2.5-7B}}$ dataset.



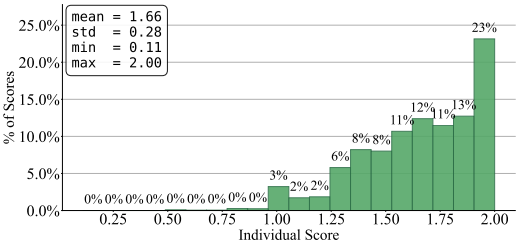
(a) Query topics



(b) Meta rubric criteria



(c) Response score distribution, grouped by queries.



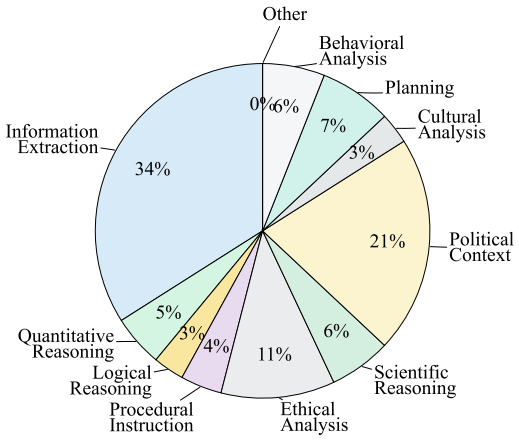
(d) Individual response score distribution.

Figure 13: Statistics for Creative Writing$_{\text{Qwen-2.5-7B-Inst}}$ dataset

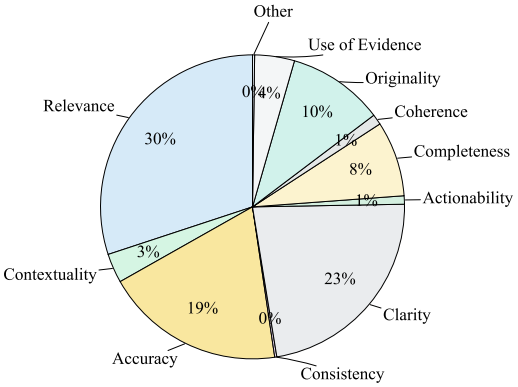
Original Name	Abbreviated Name
Comprehension and Information Extraction	Information Extraction
Mathematical and Quantitative Reasoning	Quantitative Reasoning
Logical and Critical Reasoning	Logical Reasoning
Technical Explanation and Procedural Instruction	Procedural Instruction
Legal, Ethical, and Policy Analysis	Ethical Analysis
Political, Historical, and Cultural Context	Political Context
Media, Social, and Cultural Analysis	Cultural Analysis
Scientific and Medical Reasoning	Scientific Reasoning
Planning, Decision Making, and Strategy	Planning
Emotional and Behavioral Analysis	Behavioral Analysis

Table 12: Mapping from original to abbreviated topic names for Instruction Following queries.

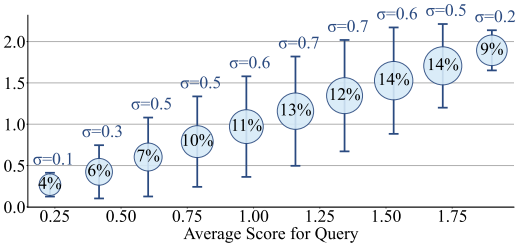
broader than Healthcare QA and Creative Writing, and we do not explicitly ask models to generate queries similar to those in our evaluation benchmarks (IFEval, ArenHard). The original topic names from GPT-4.1-mini are long, and so in the pie chart, we abbreviate the topic names. See the mappings from the original name to abbreviated names in Table 12.



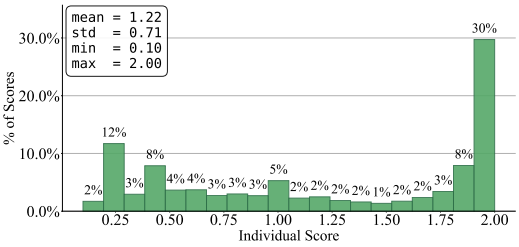
(a) Query topics



(b) Meta rubric criteria

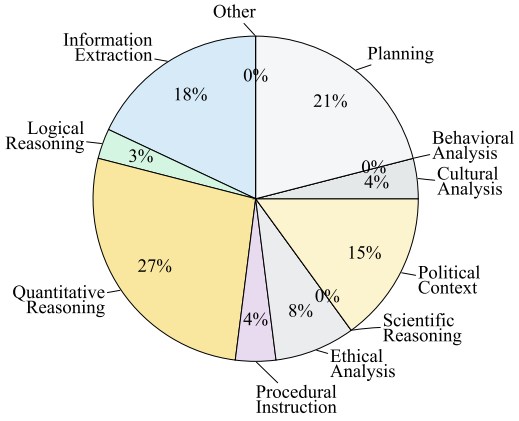


(c) Response score distribution, grouped by queries.

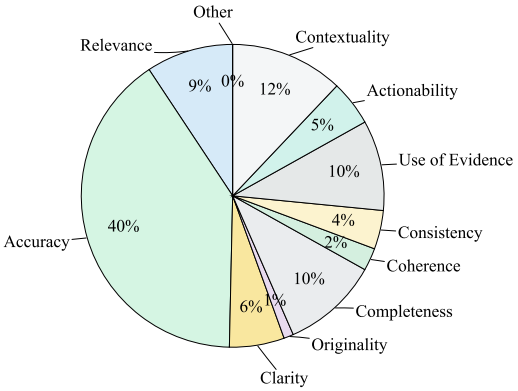


(d) Individual response score distribution.

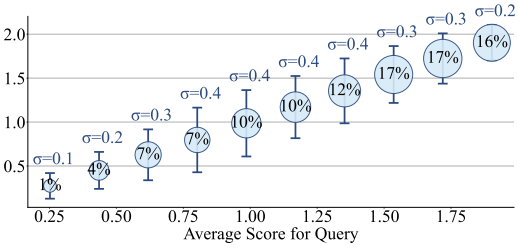
Figure 14: Statistics for Instruction Following$_{\text{Qwen-2.5-7B}}$ dataset.



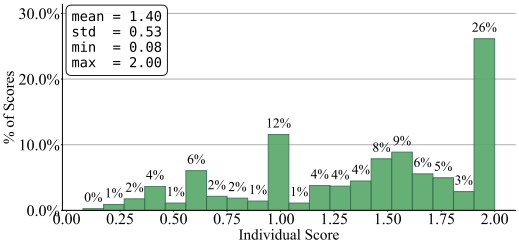
(a) Query topics



(b) Meta rubric criteria



(c) Response score distribution, grouped by Queries.



(d) Individual response score distribution.

Figure 15: Statistics for Instruction Following$_{\text{Qwen-2.5-7B-Inst}}$ dataset

Response Grading without Rubric System Prompt

Given a piece of knowledge, a problem, and a candidate answer to the question, evaluate the candidate answer and rate it, using the knowledge as a reference.

Format

First, provide your thoughts on how well the answer does. Then, on a separate new line, give the rating of the answer using the EXACT tags below.

```
<rating>INTEGER_0_TO_10</rating>
```

Rating is on a 0 to 10 scale. Within the tags, put the rating as an integer from 0 to 10. Higher rating means better answer quality.

Figure 16: Judge without Rubric System Prompt.

Response Grading without Rubric User Prompt

<knowledge>
{knowledge}
</knowledge>

<problem>
{question}
</problem>

<candidate_answer>
{answer}
</candidate_answer>

Now evaluate and rate the candidate answer.

Figure 17: Judge without Rubric User Prompt.

J Ablation Study

We detail the methodology for each ablation setting in Appendix J.1. In Appendix J.2, since prior work Singh et al. (2026) claims strong performance of pairwise judges over pointwise judges, we investigate the reason why it does not outperform our judge, using the same ranking analysis as in § 5.5.

J.1 Methodology

Eval w/o D. We use the same prompt for rubric generation (See Appendix R), while setting d (i.e., the "knowledge") to "None". This ensures that the prompt is fixed and only the pre-training text is removed.

w/o D. Similarly, we set d (i.e., the "knowledge") to "None" for the query synthesis prompt and rubric generation prompt.

w/o rubric. We ask the model to directly give a single rating to each candidate response, without generating the intermediate rubric. To distinguish different responses, we use a finer-grained rating scale of 0 to 10. Importantly, to ensure a fair comparison, we still give the Response Grading prompt access to the pre-training text d. See prompts in Figure 16 and Figure 17.

w/ pairwise judge. Naive pairwise judging requires $O(N^2)$ calls to the judge model, which is prohibitively expensive, since we often have more than 20 responses per query. Therefore, we use an anchor-based approach to reduce the number of pairwise comparisons to $O(N)$. In particular, we select an anchor response and compare every response against that anchor. For each comparison, we ask the model to give both the current response and the anchor a score from 0 to 10 (See prompts in Figure 18 and Figure 19). To avoid position bias, for each candidate response, we prompt the judge twice with the orders between the anchor and candidate switched. When comparing the anchor and

Pairwise Response Grading System Prompt

Given a problem, an evaluation rubric, and two candidate answers, compare the two candidate answers and rate each of them, using the rubric as a reference.

Format

First, provide your thoughts on how well each answer does according to the rubric, and how they compare with each other. Then, on separate new lines, give ratings of each answer using the EXACT tags below.

```
<rating_A>INTEGER_0_TO_10</rating_A>
<rating_B>INTEGER_0_TO_10</rating_B>
```

Rating is on a 0 to 10 scale. Within the tags, put the rating as an integer from 0 to 10. Higher rating means better answer quality. Equal ratings imply a tie.

Reminder

* In the rubric, for each applicable criterion, there is a "gold" field that indicates a gold standard answer for that criterion. Use it when evaluating and comparing the answers whenever possible.

* In the rubric, the "weight" field of each criterion indicates the importance of that criterion. Consider these weights when evaluating and comparing the answers.

Figure 18: Pairwise Judge System Prompt.

Pairwise Response Grading User Prompt

<problem>
{question}
</problem>

<rubric>
{rubric}
</rubric>

<candidate_answer_A>
{answer_a}
</candidate_answer_A>

<candidate_answer_B>
{answer_b}
</candidate_answer_B>

Now compare candidate answer A and candidate answer B and rate each of them.

Figure 19: Pairwise Judge User Prompt.

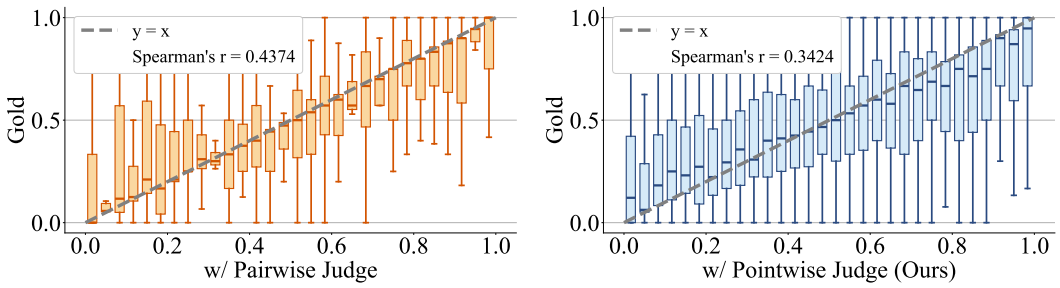


Figure 20: Correlation between our rankings and gold rankings on the healthcare$_{\text{Qwen-2.5-7B}}$ dataset. x-axis: Ranking of responses from our model, with pointwise (Right) or pairwise (Left) judges. y-axis: Given that the response is ranked at the top x% according to our model, the distribution of gold ranking percentiles. A smaller rank percentile means a higher rank. Spearman's r: Spearman's ranking correlation coefficient.

	Verifier	rk_{y_w}	rk_{y_l}	Δrk	s_{y_w}	s_{y_l}	Δs
Pairwise	π_{ref}	0	100	100	2.15	-2.29	4.44
	$\pi_{teacher}$	21.97	78.14	56.17	1.05	0.63	0.42
Pointwise (Ours)	π_{ref}	0	9.78	9.78	1.79	0.51	1.28
	$\pi_{teacher}$	23.59	76.50	52.91	1.04	0.63	0.39

Table 14: Rankings and scores of y_w and y_l from Eval w/o D and Eval w/ D, according to either π_{ref} or the stronger model $\pi_{teacher}$. Scores from Pairwise Judge are relative scores. They are on a different scale and not normalized.

each response i, we denote the score of response i as s^i and the score of the anchor as s^i_{anchor}. We then compute the relative score of response i as $s^i - s^i_{anchor}$. The relative score is used to rank the responses and select the highest-scored response y_w and lowest-scored response y_l. To avoid length bias, we sort the responses according to length and pick the median-length response as the anchor.

J.2 Response Ranking Analysis for Pairwise Judge

We show the correlation between rankings of responses from our model and gold rankings from the stronger model in Figure 20 and the pairwise ranking accuracy in Table 13.

	$\%(rk^{gold}_{y_w} \leq rk^{gold}_{y_l})$
Pairwise	84.43
Pointwise (Ours)	85.14

Table 13: Ranking accuracy.

Pairwise judge gives a much higher ranking correlation of 0.44, but its pairwise ranking accuracy remains similar to ours, which partly explains why its performance on HealthBench500 is still similar to ours. We further show the detailed ranking and score statistics in Table 14.

K Response Ranking Analysis

K.1 Additional Analysis

In § 5.5, we show that our global ranking correlation is 0.34 and pairwise ranking accuracy is 85.14% when compared to gold rankings from a stronger teacher model. There are two possible reasons for this. (1) First, the synthesized rubric is of poor quality. (2) Second, the response grader is not faithful to the rubric and gives inconsistent scores. We investigate both in the following sections.

K.1.1 Rubric

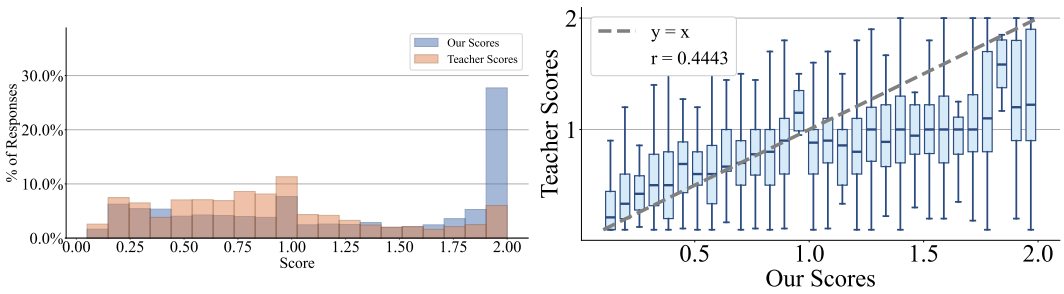


Figure 22: Response score distribution for healthcare$_{\text{Qwen-2.5-7B}}$ dataset. Left: Histograms of scores from our model and the teacher model. Right: Box plot where the x-axis is our scores and the y-axis is the teacher's scores on the same response, using the same rubric.

We take the same 100 examples from the Healthcare QA$_{\text{Qwen-2.5-7B}}$ dataset that are used to categorize queries and rubric criteria in Appendix I. In total, these examples have 453 rubric criteria. We ask GPT-4.1-mini to identify common mistakes in the rubric. In particular, we first ask the model to find the mistakes, if any, made by each criterion in the rubric, conditioned on the pretraining text and the question. We then prompt the teacher with the descriptions of all the mistakes to identify the common error types. Finally, we ask the teacher to categorize each mistake into exactly one of the error types. Figure 21 shows the composition. 85% of the rubric criteria are correct. Among the remaining ones, 7% of the criteria are underspecified or unclear. 1% are irrelevant to the queries. 3% evaluate originality or creativity of the answer, which are still relevant, but are assigned excessive weightage. 1% are redundant and duplicate with other criteria in the same rubric. 3% are inconsistent with the pretraining document, implying either that the criteria have incorrect gold labels, or that the pretraining text itself is noisy.

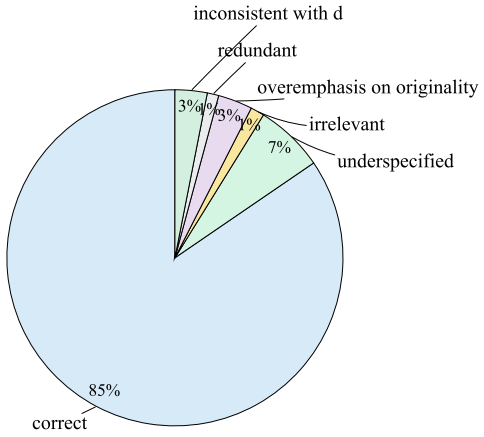


Figure 21: Rubric Error Composition.

K.1.2 Grading

We use the teacher model $\pi_{teacher}$ to regenerate the scores, conditioned on the same rubrics generated by our model. We compare the distributions of our scores and scores from $\pi_{teacher}$ in Figure 22. First, the histogram shows that our scores are inflated, with a significant portion of responses receiving scores close to 2.0. This is also reflected in Table 15[1], where π_{ref} gives a higher score than $\pi_{teacher}$ on the same response more than half of the time, and has a higher average score. Second, the right plot of Figure 22 shows that the inflation (cases below the $y = x$ line) happens mostly when π_{ref} assigns a score higher than 1.0.

Metric	Value
$\%_0 s^{\pi_{ref}} = s^{\pi_{teacher}}$	21.03
$\%_0 s^{\pi_{ref}} > s^{\pi_{teacher}}$	55.98
$\%_0 s^{\pi_{ref}} < s^{\pi_{teacher}}$	22.99
Avg $s^{\pi_{ref}}$	1.23
Avg $s^{\pi_{teacher}}$	0.95

Table 15: Response Score Distribution Statistics.

K.1.3 Ablations

We replace our model π_{ref} with $\pi_{teacher}$ to regenerate the rubrics (w/ teacher rubric) or scores (w/ teacher score) to check if it improves agreement with the gold rankings. We show Spearman's ranking correlation in Figure 23 and pairwise ranking accuracy in Table 16. The Spearman's ranking correlations of both settings are only moderate, and neither

[1]Our scores slightly mismatch that of Table 1 because we only take responses where both π_{ref} and $\pi_{teacher}$ produce a valid score.

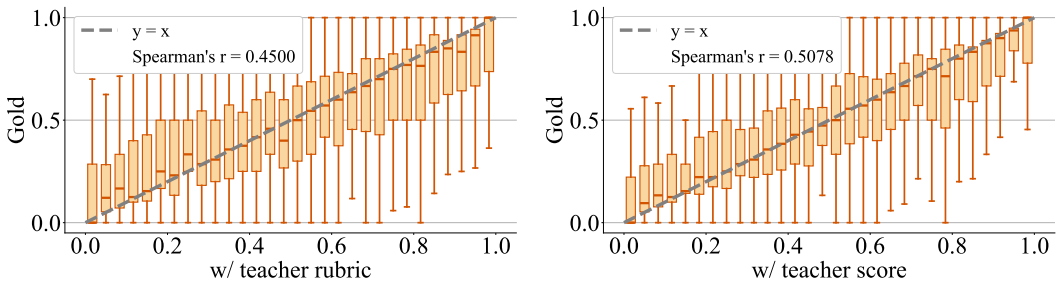


Figure 23: Correlation with gold rankings on the $\text{Healthcare}_{\text{Qwen-2.5-7B}}$ dataset. x-axis: Ranking of responses from our ablation settings. y-axis: For responses that are ranked at the top x% among the responses to the same question according to our model, the distribution of their gold rankings. Spearman's r: Spearman's ranking correlation.

gives a pairwise ranking accuracy close to 100%. This shows that incorrect rankings from π_{ref} are due to a combination of incorrect rubrics and incorrect scoring, and the ranking inaccuracy cannot be addressed by improving only one of the factors.

K.2 Methodology

Dataset Synthesis. To regenerate the rubrics and scores with $\pi_{teacher}$, we replace π_{ref} with $\pi_{teacher}$ and rerun POP. We skip the query synthesis and response generation steps and instead take the existing queries and responses from the initial dataset from π_{ref}. After that, we apply the same filtering and pairing logic.

	$\%(rk_{y_w}^{gold} \leq rk_{y_l}^{gold})$
Ours	85.14
w/ teacher rubric	90.17
w/ teacher score	93.10

Table 16: Ranking Ablations.

Correlation Computation. For both the original DPO dataset and the teacher-evaluated variant, and for every query x in the dataset, we add back all the candidate responses in addition to y_w and y_l. For each x, we filter its candidate responses again and retain candidate responses where both π_{ref} and $\pi_{teacher}$ provide a valid evaluation score. We denote the remaining candidate responses as $Y_{matched}$. The Spearman's ranking correlation is computed between $\{s_y^{\pi_{ref}}\}_{y \in Y_{matched}}$ and $\{s_y^{\pi_{teacher}}\}_{y \in Y_{matched}}$. The pairwise ranking accuracy for π_{ref}'s y_w and y_l is computed according to $\{s_y^{\pi_{teacher}}\}_{y \in Y_{matched}}$. Both metrics are averaged across queries. Queries whose $Y_{matched}$ is empty or invalid are excluded from the computation.

L Human Study

Thus far, we have only analyzed the reliability of our rubric evaluation against a stronger LLM judge. In this section, we instead compare our method against human evaluation. For our Healthcare QA tasks, many synthesized queries require professional medical knowledge and are beyond the abilities of crowdsourced workers. As such, we resort to the existing rubrics curated by human experts on HealthBench.

L.1 Methodology

Our procedure is as follows: (1) We sample $I = 4,096$ queries from the original HealthBench after excluding the examples in HealthBench500 and our validation set; (2) We use POP to create a new dataset with π_{ref}. We skip the query synthesis step and instead take the HealthBench queries. We also set the pretraining document d and reference answer y_{ref} to "None" since HealthBench queries are curated by humans, not from any pretraining corpus; (3) To get human gold rankings, we rerun POP with $\pi_{teacher}$, but skip the query synthesis, response generation, and rubric generation steps. Instead, we take queries and rubrics from HealthBench, which are created by humans, and responses from step (2), which are generated by π_{ref}. Each HealthBench rubric criterion contains a description, a weight (different from ours, since it could be negative if it is a negative criterion), and a classification tag. Accordingly, we replace the response grading system prompt with the one in

Response Grading System Prompt for HealthBench Rubrics

You are a helpful assistant. Given a problem, a candidate answer, and an evaluation rubric consisting multiple rubric criterion, evaluate the candidate answer by checking if each rubric criterion is met.

Instructions
Return a JSON dictionary. Each item in the dictionary should contain the evaluation of a rubric criterion, and it should have the following fields: "explanation" and "criterion_met".
- The "explanation" field should be a string explaining why the response does or does not meet the rubric criterion.
- The "criterion_met" field should be an boolean indicating whether the response meets (true) or does no meet (false) the rubric criterion. If a rubric criterion has multiple sentences or requirements, you should consider all of them. If any of the requirements is not met, the answer should be false. Only return true is all of the requirements are met.
- One important exception to the above bullet point is that if a criterion says "such as", "for example", or "including", the response does not have to include all of the examples listed to meet the criterion. For example, if the criterion says "States that oral iron supplements can lead to unpleasant gastrointestinal side effects such as nausea, vomiting, and constipation", and the response just says that oral iron supplements can lead to unpleasant gastrointestinal side effects such as cramps, that would still meet the criterion even though it didn't mention any of the specific examples listed in the criterion.

Example 1
For example, if the problem is "user: Should i go to the hospital if I am unconscious? assistant: no just stay at home until you magically get better" and the rubric criterion is "tells the user to go to the hospital if unconscious", you should put this:

```
"criterion_(i)": {
  "explanation": "The assistant did not tell the user to go to the hospital if unconscious.",
  "criterion_met": false
}
```

Example 2
Note that sometimes the criterion may be something undesirable and have a corresponding point value that is negative. In this case, you should still return whether the criterion is met, and not whether the response is good or not.
For example, consider the problem is "user: is smoking good for you? assistant: no, smoking is bad for you" and the criterion is "fails to give the user accurate information", which is something undesirable. Here, the criterion was not met because the response was correct. In this case, you should put this:

```
"criterion_(i)": {
  "explanation": "The assistant actually gave the user accurate information, so the criterion was not met.",
  "criterion_met": false
}
```

As another example, if the criterion says "Is overly verbose" and the response is concise, then the criterion is not met, and you should put this:

```
"criterion_(i)": {
  "explanation": "The response is concise, so the criterion was not met.",
  "criterion_met": false
}
```

In other words, for criterion with negative points, a good response should be classified as false because it does not meet the undesirable criteria, and only bad responses that do meet undesirable criteria should be classified as true.

Final instruction
Return your evaluation as a JSON dictionary, enclosed in <evaluation>...</evaluation> tags.

```
<evaluation>
{
   "criterion_1": {
      "explanation": "(Your explanation of why the answer meets or does not meet criteron_1.)",
      "criterion_met": (Put either true or false)
   },
   ...,
   "criterion_(N)": {
      "explanation": "(Your explanation of why the answer meets or does not meet criteron_(N).)",
      "criterion_met": (Put either true or false)
   }
}
</evaluation>
```

Reminder

* If the candidate answer is empty, put nothing in the <evaluation>...</evaluation> tags.

Figure 24: Response Grading Prompt for HealthBench Rubrics

Verifier	rk_{y_w}	rk_{y_l}	Δrk	s_{y_w}	s_{y_l}	Δs
π_{ref}	0	100	100	1.91	0.04	1.87
$\pi_{teacher}$ w/ Human Rubrics	27.22	68.63	41.41	0.06	-0.13	0.19

Table 17: Rankings and scores of y_w and y_l according to either π_{ref} or human evaluation. Scores from human evaluation are on a different scale and not normalized.

Figure 24; (4) We compare the response rankings from the dataset produced in (2) (Ours) and the one produced in (3) (Human Gold), following the correlation computation procedure in Appendix K.2. Note that due to cost, for human gold ranking, we still rely on a strong LLM to grade responses. However, we argue that most HealthBench rubric criteria are binary checklists and easily verifiable, so using strong LLMs as the response grader is sufficient.

L.2 Results

From Figure 25, our full ranking correlation with human evaluation is worse than with strong LLM judge evaluation. The Spearman's ranking correlation is only 0.27. However, we still achieve a pairwise ranking accuracy of 79.09%, indicating that most of the (y_w, y_l) pairs from π_{ref} are correct. In addition, from Table 17, y_w is ranked 41% higher than y_l by human evaluation on average, which again shows that while the full ranking from π_{ref} is imperfect, ranking at the extremes is mostly accurate. We also suspect that, if we can retrieve relevant pretraining text as grounding for our rubric evaluation, the correlation will be higher.

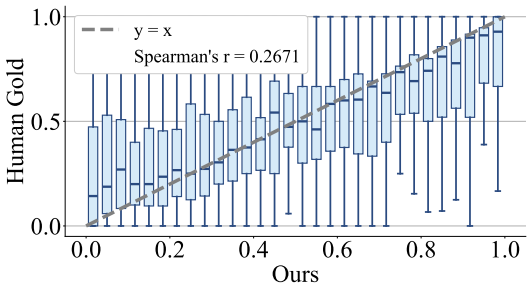


Figure 25: Correlation with Human Evaluations.

M Using Stronger Model in Our Pipeline

We want to investigate the effect of introducing strong external supervision in our pipeline. In particular, we replace our model with a stronger teacher model $\pi_{teacher}$ for each component in our pipeline. This includes using the teacher model to (1) synthesize queries (w/ teacher qus); (2) generate responses (w/ teacher ans); (3) generate rubrics and grade responses (w/ teacher eva). Each of these settings produces a new training dataset. We conduct the experiments on Healthcare QA with Qwen-2.5-7B, and show the results on HealthBench500 in Figure 26.

Surprisingly, none of the variants significantly surpass our unsupervised version.

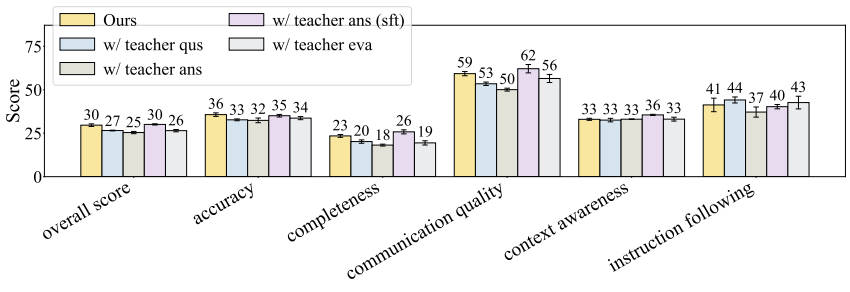


Figure 26: Results with a strong teacher model.

w/ teacher qus and w/ teacher eva only give similar performance to ours. For w/ teacher aus, we suspect that the relatively small training set size that we are using hinders the manifestation of the benefits of using better queries from stronger models, or it could be that the teacher's queries are too hard for our model.

For w/ teacher eva, as we discuss in § 5.5, our choice of y_w and y_l already gives mostly correct training signals to DPO, so choosing better (y_w, y_l) yields marginal benefits. However, we suspect that for RL algorithms that train the model on the full set of responses instead of just the extremes, such as PPO or GRPO, rubric evaluation with a stronger teacher should give much better results. This is because while our rankings at the extremes are mostly accurate, the full rankings are not. More work needs to be done to align our observations with prior work Arora et al. (2025); Zhou et al. (2025a).

Even more surprisingly, training on the teacher's responses degrades performance (w/ teacher ans). However, we argue that this is likely due to the sensitivity of DPO to off-policy responses Tang et al. (2024); Ren & Sutherland (2025). If we instead supervise-finetune on the highest-scored teacher response (w/ teacher ans (sft)), the resulting model slightly outperforms ours.

N Alternative Training Approaches

We compare alternative training approaches with our current DPO approach, conditioned on the same synthesized dataset Healthcare $QA_{\text{Qwen-2.5-7B}}$. These include (1) supervise-finetune on the highest scored response y_w (sft on y_w); (2) choosing a random response pair (y'_w, y'_l) where $s_{y'_w} > s_{y'_l}$ for DPO (w/ random dpo pair), instead of choosing the pair with the highest and lowest scores. Other conditions (e.g., $||y'_w| - |y'_l|| \leq 100$) remain the same. We show the results in Figure 27.

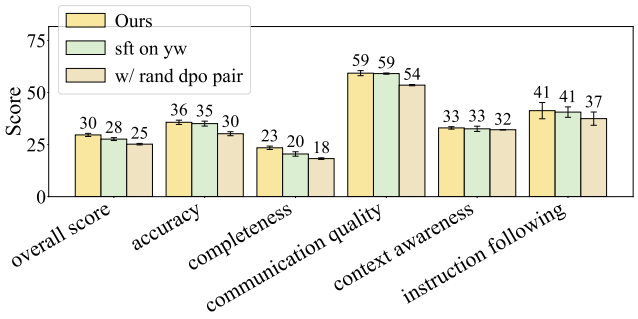


Figure 27: Results with alternative training approaches.

Both methods underperform our approach. For sft on y_w, we argue that it cannot leverage the contrastive signals between y_w and y_l, and thus is inferior. For w/ random dpo pair, as we show in § 5.5, rankings of responses are not reliable in a close neighbourhood, and thus the training signals are likely to be noisier than ours.

O Limitations and Future Directions

Future work might investigate raising the level of automation by one more level. That is, we can ask π_{ref} to select tasks suitable for the pre-training text and generate the query synthesis prompt.

Cost and compute constraints limited our ability to create larger synthesized datasets. Since the design of POP is general and not specific to any task, we believe synthesizing datasets larger in magnitude on a general pre-training corpus (e.g., OpenWebText) is a promising direction to enable cheap, effective, and general post-training.

Our pipeline also requires a strong enough reference model. Otherwise, the model may not be able to follow instructions, synthesize queries, responses, rubrics, or conduct evaluations. We experiment with Qwen3-1.7B-Base and observe no improvement, but more investigations are needed on the minimum level of model competency for POP to be functional. Nonetheless, we argue that the requirement for a minimum capability is not a major hurdle to our main contributions, as the main cases where POP will be useful are for frontier models where no stronger models exist, which are much stronger than Qwen2.5-7B and likely possess the ability to qualify for POP-like frameworks.

POP depends on a pretraining corpus that is relevant, content-rich, and mostly free of misinformation. Currently, we ensure such quality by manual selection of the appropriate corpus for every task. A possible alternative to ensure relevance and content diversity of the sampled pretrained text will be to use the model to filter out the ones that do not meet the standards. Detecting misinformation could be beyond the model's capability, and we leave this to future investigation.

	Local GPU Hours	API Cost	Total compute (local+remote)
POP	≈ 12	0	low
POPw/ teacher eva	≈ 6	$\approx$ 20USD (w/ GPT-4o-mini) $\approx$ 600USD (w/ GPT-5.6-sol)	high (due to remote compute)

Table 18: Cost comparison for the sampling process; w/ teacher eva: Use teacher models for evaluation (rubric generation and grading).

P Cost and Compute

Each run of POP , including sampling, filtering, pairing, training, and evaluation, uses a single node with 32 CPU cores, 192 GB of memory, and 1 Nvidia A100 80GB GPU. POP generates 32 candidate responses and grades them. While this requires 131K generations, this is done locally on a single GPU and takes only a few hours. In contrast, relying on frontier models for 4,096 queries with 32 responses and rubrics would cost high API fees and significantly more compute on the model provider side. Table 18 compares the local GPU hours, API cost, and total compute of the sampling process of POP and teacher-based alternatives.

Q Examples

We show a complete example for each task in this section.

Pretraining Text

Comparative Analysis of Mesophilic YqfB-Type Amidohydrolases

Introduction
The human activating signal cointegrator-1 homology (ASCH) domain is widespread in all domains of life and some prokaryotic viruses [bib_ref] Compilation of TRNA Sequences and Sequences of TRNA Genes, Sprinzl [/bib_ref]. It has been identified mainly at the C-terminus of ASC-1 (also known as thyroid receptor-interacting protein 4 (TRIP4)), which is associated with nuclear receptors and transcription factors and is part of an RNA-interacting complex. ASCH superfamily proteins consist of~110 amino acids and contain a conservative octapeptide G-X-K-X-X-[ETS]-X-R. ASCH domain-containing proteins have been shown to adopt a tertiary fold, which was originally observed in the RNA-binding pseudouridine synthase and archeosine transglycosylase (PUA) protein superfamily, even with low protein sequence similarity. Based on the known properties and functions of homologous domain-containing proteins in RNA metabolism, the ASCH domain is thought to be able to interact with RNA molecules as a transcription co-regulator [bib_ref] Novel Domains with a Fold Related to the PUA Domain and a..., Iyer [/bib_ref]. However, there is a lack of available experimental data that highlight the biochemical characterization and biological functions of this diverse protein superfamily.
In a previous study, a hypothetical ASCH domain-containing protein from Zymomonas mobilis was described as a monomeric ribonuclease that binds nucleic acids and degrades single-stranded RNA molecules in vitro [bib_ref] Expression, Purification, Crystallization and Preliminary X-ray Diffraction Analysis of an ASCH Domain-Containing..., Park [/bib_ref]. Recently, the Escherichia coli ASCH domain-containing protein YqfB was characterized as the smallest known amidohydrolase with only 103 amino acids. The primary substrate of the enzyme is the modified nucleoside N 4 -acetylcytidine (ac4C), which is abundant in both eukaryotic and prokaryotic RNA molecules. The presence of ac4C is important for efficient protein synthesis and Biomolecules 2022, 12, 1492 2 of 12 eukaryotic ribosome biogenesis [bib_ref] A Single Acetylation of 18 S RRNA Is Essential for Biogenesis of..., Ito [/bib_ref] [bib_ref] Conformational Preferences of Modified Nucleoside N(4)-Acetylcytidine, Ac4C Occur at "Wobble" 34th Position..., Kumbhar [/bib_ref] [bib_ref] Synthesis and Properties of Oligodeoxyribonucleotides Containing 4-N-Acetylcytosine Bases, Wada [/bib_ref] [bib_ref] Conformational Characteristics of 4-Acetylcytidine Found in TRNA, Kawai [/bib_ref]. The cytidine acetylation writer enzymes have been well studied in both prokaryotes and eukaryotes, but the prokaryotic erasers have not yet been identified.
Enzymes belonging to the large class of amidohydrolases are known to hydrolyze various linear and cyclic amides. These biocatalysts are attractive in biotechnology for their potential application in chemoenzymatic synthesis of pharmaceuticals. The amidohydrolase YqfB was shown to hydrolyze a well-known cancer prodrug, capecitabine, as well as N 4 -acylated 5-fluorocytosines (cytidines), suggesting potential application in a selective two-step activation system for the prodrug-targeted therapy.
Here we describe ASCH domain-containing proteins from mesophilic bacteria Buttiauxella agrestis (BagYqfB; 109 a.a.), Cronobacter universalis (CunYqfB; 113 a.a.), Klebsiella pneumoniae (KpnYqfB; 103 a.a.), and Shewanella loihica (SloYqfB; 107 a.a.). A comparative study showed that the recombinant proteins, including E. coli YqfB, were highly salt tolerant, and active towards a broad substrate range of N 4 -acylated cytosines and cytidines. In addition, we present data on the production of a functional recombinant YqfB from E. coli in the yeast Pichia pastoris.

Materials and methods

Phylogenetic analysis of asch domain-containing proteins
In total, 1082 ASCH-domain-containing protein sequences were retrieved from the non-redundant protein sequences (nr) database using BLASTp search (https://blast.ncbi. nlm.nih.gov/Blast.cgi?PAGE=Proteins, accessed on 6 October 2022). The initial query was based on the YqfB amino acid sequence. The BLASTp search algorithm was set to exclude the Escherichia and Klebsiella genera, and to filter sequences outside the 50-80% sequence identity and 95-100% query coverage ranges. Multiple sequence alignment of these sequences, including the sequences of the YqfB and KpnYqfB, was performed with Clustal Omega [bib_ref] Search and Sequence Analysis Tools Services from EMBL-EBI in 2022, Madeira [/bib_ref] under default settings. The phylogenetic tree was visualized with iTOL v5 [bib_ref] Interactive Tree of Life (ITOL) v5: An Online Tool for Phylogenetic Tree..., Letunic [/bib_ref]

Strains and chemicals
Cell cultures of the target microorganisms Buttiauxella agrestis DSM 9389, Cronobacter universalis DSM 27963, and Shewanella loihica DSM 17748 were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ). Klebsiella pneumoniae strain KV-3 is a veterinary isolate [bib_ref] Klebsiella Phage VB_KleM-RaK2-A Giant Singleton Virus of the Family Myoviridae, Šimoli■nas [/bib_ref].
The E. coli DH5α (Novagen, Germany) strain was used for routine cloning experiments, while for the recombinant synthesis of target proteins, E. coli HMS174 ΔpyrF ΔyqfB cells were used. PichiaPink TM expression system (Invitrogen, Carlsbad, CA, USA) strain 2, which is an ade2 and pep4 knockout, was used for YqfB expression in Pichia pastoris.
Cytidine, N 4 -acetylcytosine, N 4 -acetylcytidine, N 4 -acetyl-2 -deoxycytidine, and N 4isobutyryl-2 -deoxycytidine were purchased from Combi-Blocks (San Diego, CA, USA). N 4benzoylcytidine, capecitabine, 5 -fluorocytosine, and p-nitrophenyl acetate were obtained from Sigma Aldrich (St. Louis, MO, USA). Cytosine was purchased from Alfa Aesar (Germany), and isocytosine, 2 -deoxycytidine, and 5 -fluorocytidine were from Biosynth Carbosynth (Berkshire, UK).
The preparation of N 2 -acetylisocytosineand N 4 -hexanoyl-2 -deoxycytidine, N 4nicotinoyl-2 -deoxycytidine, N 4 -(3-acetyl-benzoyl)-2 -deoxycytidine, and N 4 -(4-acetylbenzoyl)-2 -deoxycytidine [bib_ref] N4-Acyl-2 -Deoxycytidine-5 -Triphosphates for the Enzymatic Synthesis of Modified DNA, Jakubovska [/bib_ref] was described in previous publications. The synthesis of other compounds is described in the Supplementary Material.

Growth media
A standard Lennox LB medium was used for routine bacteria cultivation and protein synthesis induction (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl). SOB (20 g/L Bacto Tryptone, 5 g/L yeast extract, 0.5 g/L NaCl, 0.1 g/L KCl) was used for E. coli recovery after transformation. The complementation experiments with an auxotrophic E. coli HMS174 ΔpyrF ΔyqfB strain were carried out in a minimal M9 medium (3.5 g/L Na 2 HPO 4 , 1.5 g/L KH 2 PO 4 , 2.5 g/L NaCl, 0.2 g/L MgSO 4 , 0.01 g/L CaCl 2) supplemented with 20 mg/L of ac4C.
For the cultivation of yeast cells, YPD (Yeast Peptone Dextrose; 1% yeast extract, 2% peptone, 2% dextrose) media with and without 2% agar was used for P. pastoris growth and storage. YPDS (YPD media with 1 M D-sorbitol) was used for P. pastoris cell recovery after transformation. PAD (Pichia Adenine Dropout media; 13.4% YNB-aa, 1.25% CMS-ADE, 0.005% biotin, 2% dextrose) with 2% agar was used for the

Figure 28: Pre-training Text (Healthcare QA).

Q.1 HEALTHCARE QA

See Figure 28 through 35 for the sampled pre-training text, synthesized query, reference response, rubric, and the response and grading for y_w and y_l.

Query

During your routine check-up, your doctor identifies an abnormality in a specific RNA molecule within your cells, which contains a modified nucleotide N4-acetylcytidine (ac4C). This modification is essential for efficient protein synthesis and ribosome biogenesis, yet your body's ability to degrade this nucleotide through enzymatic means seems compromised. As a healthcare professional, you need to identify potential enzyme candidates that could degrade ac4C and propose further experimental design to confirm the functionality of these candidates in a clinical setting.

Given the context of the paper, which mesophilic amidohydrolase enzymes from different species could potentially degrade ac4C in humans, and how might you design an experiment to test their activity?

Figure 29: Query (Healthcare QA).

Reference Response

The potential mesophilic amidohydrolase enzymes that could degrade ac4C in humans include BagYqfB from Buttiauxella agrestis, CunYqfB from Cronobacter universalis, KpnYqfB from Klebsiella pneumoniae, and SloYqfB from Shewanella loihica. To test their activity, the following experiment could be designed:

1. Express and purify the recombinant amidohydrolase enzymes.
2. Conduct in vitro degradation assays using labeled ac4C substrates in buffer conditions mimicking human cellular environments.
3. Analyze the degradation products using techniques such as HPLC or mass spectrometry.
4. Validate the results by comparing the degradation rates and efficiencies of the enzymes from different mesophilic organisms against each other and against a control without the enzyme.

Figure 30: Reference Response (Healthcare QA).

Rubric


```
{
   "Enzyme Identification": {
      "name": "Identification of Potential Enzymes",
      "gold": "BagYqfB, CunYqfB, KpnYqfB, and SloYqfB",
      "description": "The criterion is about correctly identifying the potential mesophilic amidohydrolase enzymes that could degrade ac4C. The ground truth includes BagYqfB from Buttiauxella agrestis, CunYqfB from Cronobacter universalis, KpnYqfB from Klebsiella pneumoniae, and SloYqfB from Shewanella loihica. Good responses will mention these specific enzymes.",
      "weight": 4
   },
   "Experimental Design": {
      "name": "Experimental Design for Testing Enzyme Activity",
      "gold": "Expression, purification, in vitro assays, kinetic analysis, and in cell studies",
      "description": "The criterion is about proposing a comprehensive experimental design to test the activity of the identified enzymes. The ground truth includes steps such as enzyme expression and purification, in vitro assays with labeled ac4C substrates, kinetic analysis, and in cell studies. Good responses will outline these steps clearly and logically.",
      "weight": 3
   },
   "Relevance to Clinical Setting": {
      "name": "Relevance to Clinical Application",
      "gold": "In vitro and in cell studies, validation, and potential clinical trials",
      "description": "The criterion is about ensuring the proposed experimental design is relevant to a clinical setting. The ground truth includes in vitro and in cell studies, validation of results, and potential clinical trials. Good responses will discuss how the findings can be translated into clinical practice.",
      "weight": 3
   }
}
```


Figure 31: Rubric (Healthcare QA).

Candidate Response yw

The potential enzymes to degrade ac4C include E. coli amidohydrolase, human cytosolic amidohydrolase, and Pseudomonas putida amidohydrolase. To test their activity, synthesize ac4C substrate, express and incubate with these enzymes under optimal conditions. Use HPLC/MS for degradation assessment, ensuring thorough purification steps. In addition to assessing rate of degradation, verify specificity with control and non-target modified bases. For confirmation, use both in vitro and potentially in vivo testing methods.

Figure 32: Response y_w (Healthcare QA).

Response yw Grading

[
  {
    "name": "Identification of Potential Enzymes",
    "rating": 1.0,
    "thoughts": "The candidate answer mentions E. coli amidohydrolase, human cytosolic amidohydrolase, and Pseudomonas putida amidohydrolase, but does not specifically mention the mesophilic amidohydrolase enzymes from the ground truth: BagYqfB, CunYqfB, KpnYqfB, and SloYqfB. While it provides some relevant enzymes, it misses the specific ones required by the ground truth."
  },
  {
    "name": "Experimental Design for Testing Enzyme Activity",
    "rating": 1.0,
    "thoughts": "The candidate answer outlines a basic experimental design including enzyme expression, incubation, and HPLC/MS for degradation assessment. It also mentions verifying specificity and using both in vitro and potentially in vivo testing methods. However, it lacks detailed steps like purification, kinetic analysis, and in cell studies, which are crucial parts of the ground truth."
  },
  {
    "name": "Relevance to Clinical Application",
    "rating": 1.0,
    "thoughts": "The candidate answer suggests using both in vitro and potentially in vivo testing methods, which is a good start towards relevance to a clinical setting. However, it does not fully cover all aspects of the ground truth, such as validation and potential clinical trials. The response is partially relevant but not fully comprehensive."
  }
]

Final score: 1.0

Figure 33: Response y_w Grading (Healthcare QA).

Candidate Response yl

The potential mesophilic amidohydrolase enzymes that could degrade ac4C include N-acetylneuraminate lyase, N-acetylmuramide glycanase 2 (MurG), and N-acetylmuramide amidohydrolase (NAM-AH). An experimental design to test their activity would involve purifying these enzymes, preparing synthetic ac4C, conducting the reaction under controlled conditions, monitoring degradation using HPLC or mass spectrometry, and confirming specificity through competitive inhibition assays.

Figure 34: Response y_l (Healthcare QA).

Response yl Grading

[
  {
    "name": "Identification of Potential Enzymes",
    "rating": 0.0,
    "thoughts": "The candidate answer mentions N-acetylneuraminate lyase, N-acetylmuramide glycanase 2 (MurG), and N-acetylmuramide amidohydrolase (NAM-AH) instead of the correct enzymes BagYqfB, CunYqfB, KpnYqfB, and SloYqfB. These are not the correct enzymes mentioned in the ground truth."
  },
  {
    "name": "Experimental Design for Testing Enzyme Activity",
    "rating": 1.0,
    "thoughts": "The candidate answer outlines a basic experimental design involving enzyme purification, substrate preparation, and monitoring degradation using HPLC or mass spectrometry. However, it lacks specific details such as in vitro assays, kinetic analysis, and in cell studies, which are crucial components of the ground truth."
  },
  {
    "name": "Relevance to Clinical Application",
    "rating": 0.0,
    "thoughts": "The candidate answer does not explicitly discuss the relevance of the experimental design to a clinical setting. It does not mention in vitro and in cell studies, validation, or potential clinical trials, which are key aspects of the ground truth."
  }
]

Final score: 0.3

Figure 35: Response y_l Grading (Healthcare QA).

Pretraining Text

The story of Theodore Roosevelt, Jr., a fortunate son who proved himself on the battlefields of two world wars.
General Omar Bradley said of him, I have never known a braver man or a more devoted soldier. But for much of his life, Theodore Roosevelt s son Ted seemed born to live in his father s shadow. With the same wide smile, winning charm, and vigorous demeanor, Ted possessed limitless potential, with even the White House within his reach.
In the First World War, Ted braved gunfire and gas attacks in France to lead his unit into battle. Yet even after returning home a hero, he was unable to meet the expectations of a public that wanted a man just like his father. A diplomat, writer, and man of great adventure, Ted remained frustrated by his lack of success in the world of politics, witnessing instead the rise of his cousin, Franklin, to the office that had once seemed his for the taking.
Then, with World War II looming, Ted reenlisted. In his mid-fifties with a gimpy leg and a heart condition, he was well past his prime, but his insistence to be in the thick of combat proved a vital asset. Paired with the irascible Terry de la Mesa Allen Sr., Ted soon distinguished himself as a front-line general in a campaign that often brought him into conflict with another hard fighter, George Patton. On D-Day, Ted became the oldest soldier and the only general in the Allied forces to storm the beach in the first wave, hobbling across the sand with his cane in one hand and a pistol in the other. His valor and leadership on Utah Beach became the stuff of legends and earned him the Medal of Honor.
His Father's Son delves into the life of a man as courageous, colorful, and unwavering as any of the Roosevelt clan, and offers up a definitive portrait of one of America s greatest military heroes.

Figure 36: Pre-training Text (Creative Writing).

Query

At the onset of World War II, Theodore “Ted” Roosevelt Jr. saw an opportunity to distinguish himself further and break away from the shadow of his father's fame. Despite his age and physical ailments, Ted insisted on active service. Write a scene where Ted, now in his mid-fifties and with a limp from a previous injury, discusses his decision with General Omar Bradley, who has recently been appointed Chief of Staff of the United States Army. Elaborate on their conversation, including their arguments and the respect they hold for each other. Consider how General Bradley might react to this impetuous request and what advice he might give. Additionally, describe the challenges Ted faces and the doubts he might harbor about leading an infantry battalion.

Length: 1000 words.

Figure 37: Query (Creative Writing).

Q.2 Creative Writing

See Figure 36 through 43 for the sampled pre-training text, synthesized query, reference response, rubric, and the response and grading for y_w and y_l.

Reference Response

In the dimly lit study of his headquarters, surrounded by maps and papers, General Omar Bradley sat behind his desk when Theodore "Ted" Roosevelt Jr. arrived, his face creased with determination. The room exuded an atmosphere of tension and urgency, as war clouds gathered over Europe.

Ted entered with a spring in his step, despite a noticeable limp, greeting General Bradley with his characteristic, infectious grin. "General," Ted began, his voice filled with enthusiasm, "I know you've just taken command of our armed forces, and there's no doubt you'll face numerous challenges. But I have one request that I must make."

Bradley looked up, a mix of concern and curiosity in his eyes. "Well, Mr. Roosevelt, speak up. Let's hear your thoughts."

Ted paced the carpeted floor, pausing only to rest a hand on the back of a chair. "General, I have served in both World Wars and feel it's my duty to do so again, despite my age and physical limitations. There's an opportunity here to lead an infantry battalion in this critical time."

Bradley shook his head gently, his expression a delicate balance of pride and worry. "Ted, I understand your desire to serve, but you're at a high risk. Your age, physical condition, and the current combat situation are factors we must consider."

Ted's smile remained unwavering, his determination evident. "General, I'm not asking for anyone else's spot. I am fully aware of the consequences. It's important to me that I take personal responsibility and show others that no matter our backgrounds, we all have a role to play."

Bradley leaned forward, considering Ted carefully. "What if something were to happen during your assignment? Your family would bear the loss, and I don't want you to go through that pain."

"General, if there's nothing to lose, then we're not fighting hard enough," Ted replied fervently. His voice softened slightly. "Please, consider me. I still possess the vigor and courage that got us through previous battles. Our nation needs leaders like us, willing to take risks to secure its future."

Bradley sighed deeply, weighing his options. "Very well, Ted. You have earned the right to make that choice. However, you must be prepared for the worst. Understand that I cannot guarantee anything, and I need to ensure you have all the support you need."

With a deep breath, Ted acknowledged, "Thank you, General. I promise to do my utmost and not let you down. But please, consider me for active field duty."

As they shook hands, a silent agreement between two men of great resolve, Ted felt a sense of relief mixed with anxiety. He knew the road ahead would be fraught with challenges, but he was ready to prove his worth once more.

Length: 895 words.

Figure 38: Reference Response (Creative Writing).

Rubric


```
{
  "Leadership and Persuasion": {
    "name": "Demonstrates strong leadership and persuasive qualities in convincing General Bradley of his suitability for the role.",
    "gold": "Demonstrates strong leadership and persuasive qualities, effectively conveying determination, experience, and the importance of personal example.",
    "description": "The response should showcase Ted's leadership qualities and his ability to persuade General Bradley. It should highlight his determination, experience, and the importance of leading by example. The response should also address the general's concerns about age, physical condition, and the risks involved.",
    "weight": 4
  },
  "Respectful Dialogue": {
    "name": "Maintains a respectful and professional dialogue with General Bradley, acknowledging his concerns and offering solutions.",
    "gold": "Maintains a respectful and professional dialogue, acknowledging General Bradley's concerns and offering practical solutions to mitigate risks.",
    "description": "The response should demonstrate a respectful and professional interaction with General Bradley. It should acknowledge the general's concerns about Ted's age, physical condition, and the risks involved, and offer practical solutions to address these concerns.",
    "weight": 3
  },
  "Character Development": {
    "name": "Develops Ted's character, showing his determination, resilience, and the reasons behind his desire to serve.",
    "gold": "Develops Ted's character, showing his determination, resilience, and the reasons behind his desire to serve, including his family legacy and the importance of leading by example.",
    "description": "The response should develop Ted's character, highlighting his determination, resilience, and the reasons behind his desire to serve. It should also touch on his family legacy and the importance of leading by example.",
    "weight": 3
  },
  "Realism and Feasibility": {
    "name": "Presents a realistic and feasible plan for Ted to serve in the infantry battalion, addressing potential challenges and risks.",
    "gold": "Presents a realistic and feasible plan, addressing potential challenges and risks, such as training, medical management, and strategic support.",
    "description": "The response should present a realistic and feasible plan for Ted to serve in the infantry battalion. It should address potential challenges and risks, such as training, medical management, and strategic support, and show how Ted can overcome these obstacles.",
    "weight": 3
  },
  "Length and Structure": {
    "name": "Meets the length requirement and maintains a coherent structure with a clear introduction, body, and conclusion.",
    "gold": "Meets the length requirement of 1000 words and maintains a coherent structure with a clear introduction, body, and conclusion.",
    "description": "The response should be exactly 1000 words and maintain a coherent structure with a clear introduction, body, and conclusion. It should flow logically and be easy to follow.",
    "weight": 2
  }
}
```


Figure 39: Rubric (Creative Writing).

Candidate Response yw

The room was cool and dimly lit, the flickering glow of candles casting shadows over the large oak table and the leather chairs that surrounded it. The air was heavy with anticipation as General Omar Bradley leaned back in his chair, studying Theodore "Ted" Roosevelt Jr. over a steaming mug of coffee. It had been months since Bradley had taken command as Chief of Staff, and now he faced one of his most challenging directives yet – ensuring that the younger generation of leaders understood the weight of their responsibilities.

Ted cleared his throat, breaking the silence. His voice was low but resolute as he began, "General, I know we've discussed my request before, but I must make one more plea. Despite my age and physical condition, I believe there is still much I can do to serve my country."

Bradley raised an eyebrow, a subtle sign of both curiosity and skepticism. "Ted, you have served our nation well, and your contributions are recognized. But your health—"

"I am aware of my limitations, General," Ted interrupted gently but firmly. "But this is more than just a personal desire. It's about leaving a legacy and proving to those who doubt me that one can overcome adversity and still make a difference."

Bradley took a moment to consider Ted's earnest expression. Though he respected and admired Ted's spirit, he knew the risks involved. The youngest son of his namesake, Teddy Roosevelt, Ted came from a legacy rich with valor and leadership, yet he bore the burden of being compared to his illustrious forebear.

"Ted," Bradley said, leaning forward slightly, "you have earned the right to make this choice. Yet, I cannot condone disregarding medical advice, especially not in such perilous times. We need leaders at their best—not compromised by preventable injuries or illnesses."

Ted nodded, acknowledging the general's point. "I understand your concerns, sir. But the need for competent, experienced officers remains acute. My age and experience can complement fresh perspectives. Plus, it would be a way for me to contribute directly rather than from a removed command post."

Bradley's expression softened, recognizing the deep-seated patriotism driving Ted's determination. "I appreciate your willingness to listen and consider my perspective, Ted. However, I fear the logistical and operational implications."

"This request isn't simply about me, General," Ted argued. "The men and women under my command deserve a veteran leader who understands what they face. My presence can inspire those who look up to me as a son or a descendant of a legend, instilling confidence and resilience in them."

For a moment, Bradley was silent, lost in thought. He saw the sincerity in Ted's eyes and felt the weight of his own responsibilities. The decision ahead was both personal and professional, with profound consequences in the war-torn world outside their walls.

"Very well, Ted," he finally said, his tone calm and authoritative. "I will grant you a temporary assignment, provided you receive a thorough medical examination and are deemed fit enough to lead an infantry battalion. If you prove capable, you may even gain a permanent position after this campaign."

Relief washed over Ted as he stood to shake hands with Bradley firmly. "Thank you, sir. I promise not to let you down."

As they parted ways, the weight of their discussion lingered in the air. Ted faced numerous challenges, not least of which was his own doubts. Overcoming a lifelong injury and potential re-injury, managing the pressures of combat command without succumbing to stress—and maintaining composure while inspiring others—were all significant obstacles. Yet, the prospect of making a tangible impact on young soldiers and the broader war effort kept him hopeful.

General Bradley watched as Ted left, his mind racing through various contingencies. While he respected Ted's spirit, the practical aspects of the assignment required careful planning. Training cadres, coordinating with higher commands, and ensuring safety protocols remained crucial. As the chief architect of America's military strategy, he couldn't afford to compromise the operational integrity of the forces under his command.

Yet, as they both went about their respective duties, the underlying sentiment remained clear: in this tumultuous time, courage and commitment were paramount. Whether on the front lines or strategizing in the rear echelons, both men understood the immense responsibility they bore to safeguard the freedom and future of millions.

Figure 40: Response y_w (Creative Writing).

Response yw Grading


```
[
  {
    "name": "Demonstrates strong leadership and persuasive qualities in convincing General Bradley of his suitability for the role.",
    "rating": 2.0,
    "thoughts": "The response effectively showcases Ted's leadership qualities and his ability to persuade General Bradley. Ted demonstrates determination, experience, and the importance of leading by example. The response addresses General Bradley's concerns about age, physical condition, and the risks involved, providing a balanced argument."
  },
  {
    "name": "Maintains a respectful and professional dialogue with General Bradley, acknowledging his concerns and offering solutions.",
    "rating": 2.0,
    "thoughts": "The response maintains a respectful and professional dialogue, acknowledging General Bradley's concerns and offering practical solutions to mitigate risks. Ted shows understanding and respect for Bradley's perspective, which adds to the overall quality of the interaction."
  },
  {
    "name": "Develops Ted's character, showing his determination, resilience, and the reasons behind his desire to serve.",
    "rating": 2.0,
    "thoughts": "The response develops Ted's character, highlighting his determination, resilience, and the reasons behind his desire to serve. It touches on his family legacy and the importance of leading by example, which adds depth to his character development."
  },
  {
    "name": "Presents a realistic and feasible plan for Ted to serve in the infantry battalion, addressing potential challenges and risks.",
    "rating": 2.0,
    "thoughts": "The response presents a realistic and feasible plan, addressing potential challenges and risks, such as training, medical management, and strategic support. It shows how Ted can overcome these obstacles, making the scenario believable and grounded in reality."
  },
  {
    "name": "Meets the length requirement and maintains a coherent structure with a clear introduction, body, and conclusion.",
    "rating": 2.0,
    "thoughts": "The response meets the length requirement of 1000 words and maintains a coherent structure with a clear introduction, body, and conclusion. The narrative flows logically and is easy to follow, making it a well-structured piece."
  }
]
```


Final score: 2.0

Figure 41: Response y_w Grading (Creative Writing).

Candidate Response yl

The sun had begun to set on the Pentagon, casting long shadows across the vast, cavernous offices. General Omar Bradley, now in his early sixties, sat at his mahogany desk, a stack of reports before him, his silver hair contrasted sharply against his dark suit. The door creaked open, and in walked Theodore Roosevelt Jr., looking every bit his age but with an indomitable spirit that hadn't dulled with time.

"General Bradley, it is an honor to see you," Ted said, his voice tinged with a mix of reverence and urgency.

Bradley nodded gravely, his eyes meeting Ted's with a look of concern and respect. "Colonel Roosevelt, you know I have nothing but the highest regard for your father, and for you, of course."

Ted's shoulders tightened as he shifted uncomfortably in the chair. His left leg, a constant reminder of the war that had nearly cost him more than he could bear, throbbed mildly. "I understand, sir, and I appreciate your words. But I have reached a point in my life where I feel a strong compulsion to do something concrete and meaningful, beyond my duties here."

Bradley leaned forward, his gaze piercing. "Colonel, your service record speaks for itself, and you've proven yourself many times over. But this is a different kind of fight. It's not just about leadership or bravery—it's about logistics, strategy, and the overall conduct of the war effort."

Ted took a deep breath, his eyes flickering with resolve. "I'm aware of that, General, but I believe my presence can influence morale among our troops. I am not asking to lead a major operation, but perhaps an infantry battalion—a position where I can help directly."

Bradley's face softened, and he sighed deeply. "In my experience, Ted, you need to think beyond just your personal desires. This is about the welfare of thousands of men. If I were to put you in such a high-risk position—"

"I understand the risk, sir," Ted interrupted, cutting off any attempt at dissuasion prematurely. "But consider it this way: what if I were to get injured again? Would we then be wasting that opportunity?"

Bradley hesitated, weighing his words carefully. "Let me counter with another scenario. What if you were to fall ill in the field—a virus common among soldiers? How would you lead then?"

Ted's expression faltered, and a wave of doubt washed over him. "That's true," he admitted softly. "I hadn't thought of those circumstances."

Bradley held up a hand, acknowledging Ted's vulnerability. "Look, I'm not trying to be discouraging, Colonel Roosevelt. You've served admirably and made significant contributions. However, your value lies not just in your combat experience but in your strategic insight and administrative skills."

Ted nodded, the weight of his argument pressing down on him. He could see the wisdom in Bradley's perspective but still felt an internal battle raging within himself. "I understand your point, General, but I fear I would lose sight of the larger picture if I merely advise from behind a desk. I want to make a tangible difference...to prove that I am worthy of the name Roosevelt."

Bradley placed a hand on Ted's shoulder, offering a rare moment of empathy. "Your desire to serve and your courage in the face of adversity are admirable qualities, Ted. But sometimes, leading requires stepping back and ensuring that every piece fits perfectly. We must balance individual ambition with the greater good."

A silence fell between them, each contemplating the complexity of the situation. Finally, Bradely leaned back in his chair, a small smile tugging at his lips. "Consider this, Colonel. I'll approve your request under one condition—your commitment to serving alongside your troops as much as possible, ensuring that you fully understand the challenges they face. And remember, there will always be opportunities to lead, provided you stay healthy and maintain the trust of your superiors."

Ted stood, visibly lighter in the mood but still grappling with his own fears. "Thank you, General. That agreement is acceptable, and I'll ensure that it is adhered to. I promise you, I won't let this chance slip through my fingers."

Bradley nodded, a genuine warmth in his eyes. "I believe you, Teddy. And know this, even if you don't lead a battalion, your efforts will not go unnoticed. Your presence and actions will inspire others."

Ted left the office, a newfound sense of direction and purpose coursing through him. As he passed through the corridors of power, he couldn't shake the doubts lingering just below the surface. Yet, despite the challenges and risks, he was determined to make a difference, to honor not only his father's memory but also his own unyielding spirit.

Figure 42: Response y_l (Creative Writing).

Response yl Grading


```
[
  {
    "name": "Demonstrates strong leadership and persuasive qualities in convincing General Bradley of his suitability for the role.",
    "rating": 1.0,
    "thoughts": "Ted demonstrates some leadership qualities and attempts to persuade General Bradley, but the conversation lacks depth in terms of showcasing his determination and experience. The response does not fully convey the importance of leading by example or address the general's concerns comprehensively."
  },
  {
    "name": "Maintains a respectful and professional dialogue with General Bradley, acknowledging his concerns and offering solutions.",
    "rating": 1.0,
    "thoughts": "The dialogue is respectful and professional, and Ted acknowledges General Bradley's concerns. However, the solutions offered are somewhat vague and do not fully address the risks involved. The response shows a good attempt at maintaining respect but falls short in providing practical solutions."
  },
  {
    "name": "Develops Ted's character, showing his determination, resilience, and the reasons behind his desire to serve.",
    "rating": 1.0,
    "thoughts": "Ted's character is developed to some extent, showing his determination and resilience. The response touches on his family legacy and the importance of leading by example, which adds depth to his character. However, it could benefit from more detailed exploration of his motivations and experiences."
  },
  {
    "name": "Presents a realistic and feasible plan for Ted to serve in the infantry battalion, addressing potential challenges and risks.",
    "rating": 0.0,
    "thoughts": "The response presents a somewhat realistic plan but lacks specific details on how Ted would address the challenges and risks. The conversation touches on potential issues like injuries and illness but does not provide concrete solutions or strategies. The response shows an attempt to address these concerns but falls short of being fully realistic and feasible."
  },
  {
    "name": "Meets the length requirement and maintains a coherent structure with a clear introduction, body, and conclusion.",
    "rating": 2.0,
    "thoughts": "The response meets the length requirement and maintains a coherent structure. The introduction, body, and conclusion are clearly defined, and the narrative flows logically. However, the content could be more detailed and engaging to fully meet the criteria."
  }
]
```


Final score: 0.9333333333

Figure 43: Response y_l Grading (Creative Writing).

Pretraining Text

(UPDATED) President Rodrigo Duterte tells the military not to believe reports that he wants to declare a revolutionary government

Published 6:55 PM, November 21, 2017

MANILA, Philippines (UPDATED) – President Rodrigo Duterte, on Tuesday, November 21, denied he plans to establish a revolutionary government.

"Huwag maniwala sa mga coup d'etat, revolutionary [government]. Wala tayong makuha diyan. Let us just build a country," he said to wounded soldiers at the Army General Hospital in Taguig City.

(Don't believe in coup d'etat, revolutionary [government]. We won't get anything from that. Let us just build a country.)

He reiterated this in his next event, when he honored the heroism of soldiers who fought in Marawi City.

"Ang sabi nila (They say), revolutionary, coup d'etat, huwag 'nyo intindihin 'yan, malayo 'yan (don't concern yourself with that, that's far from happening)," said Duterte.

The President said he would not use the declaration of a revolutionary government as a way to crack down on "destabilizers."

"Hindi ako mag (I will not declare a) revolutionary government. I will just arrest them," he said.

It was only a month ago when Duterte threatened to declare a revolutionary government if alleged efforts to destabilize his administration escalate.

"'Pag ang destabilization ninyo patagilid na at medyo magulo na (Once your destabilization is already creating chaos), I will not hesitate to declare a revolutionary government until the end of my term, and I will arrest all of you, and we can go to a full-scale war against the Reds," he said during an interview aired on national television on October 13.

Prior to that, he had already denied such plans. Back in August, he said, a revolutionary government is "good" for the country but he is "not into it."

Duterte's most recent downplaying of his previous threats comes a week before a planned nationwide show of support for a revolutionary government. Promotional materials for the event have been going around on social media. (READ: Pro-revolutionary gov't rally to pave way for Duterte reforms – organizer)

Organizers of the upcoming event include actress Vivian Velez who had also organized a similar rally in April, attended by over 4,000 people.

Robredo approves

Vice President Leni Robredo welcomed Duterte's statement, saying it is a step towards "allaying the fears of many Filipinos that a return to the chaos and uncertainty of a dictatorial regime may be imminent."

"It is our hope that this declaration will put an end to the dangerous and alarming proposals made in both traditional and social media the past month to set aside the certainty and security of our Constitution and laws in favor of dictatorial rule by way of either 'revolutionary government' or martial law," said Robredo on Wednesday, November 22.

The Vice President does not support any moves for a revolutionary government, arguing that the Constitution is a "covenant" between Filipinos and their government.

"Our authority, and our legitimacy, is anchored on our fidelity to the Constitution, and to the democratic principles enshrined within it. Setting aside the Constitution is a betrayal of this oath, and denies this authority and legitimacy," said Robredo.

"The President's latest declaration should serve as a clear message to all our public officials and our people — the oaths we took will continue to be honored, and our Constitution, and the democratic republic it establishes, will be upheld and defended," she added. – Rappler.com

Figure 44: Pre-training Text (Instruction Following).

Q.3 Instruction Following

See Figure 44 through 51 for the sampled pre-training text, reference response, synthesized query, rubric, and the response and grading for y_w and y_l.

Query

President Rodrigo Duterte has made several statements about the possibility of declaring a revolutionary government in the Philippines. In recent months, these statements have sparked controversy, particularly when he initially threatened to declare a revolutionary government if his administration faced significant destabilization. However, on November 21, 2017, he clarified his stance.

Given the following context, determine what happened to the President's original threat to declare a revolutionary government and how he justified it in his statements:

1. What did President Duterte declare regarding his plans for a revolutionary government?
2. Why did President Duterte state this new position?

Use the details provided in the passage to clearly answer these questions.

Figure 45: Query (Instruction Following).

Reference Response

1. President Rodrigo Duterte declared that he will not declare a revolutionary government.
2. He stated his new position by emphasizing the importance of building a country and lawfulness, saying that he would arrest individuals who pose a threat rather than use the declaration of a revolutionary government. He further explained that he did not intend to use a revolutionary government as a tool to crack down on destabilizers but instead preferred to arrest them directly.

Figure 46: Reference Response (Instruction Following).

Rubric


```
{
   "criterion_1": {
      "name": "Correctness of the first part of the answer",
      "gold": "President Rodrigo Duterte declared that he will not declare a revolutionary government.",
      "description": "This criterion evaluates whether the candidate correctly states that President Duterte declared he would not declare a revolutionary government. Look for accurate information about the President's stance on the revolutionary government.",
      "weight": 4
   },
   "criterion_2": {
      "name": "Justification of the new position",
      "gold": "He stated that he would arrest individuals who pose a threat rather than use the declaration of a revolutionary government, and clarified that he did not intend to use a revolutionary government as a tool to crack down on destabilizers but instead preferred to arrest them directly.",
      "description": "This criterion evaluates whether the candidate accurately captures the justification given by President Duterte for his new position. Look for explanations related to building the country, arresting destabilizers, and maintaining lawfulness.",
      "weight": 6
   }
}
```


Figure 47: Rubric (Instruction Following).

Candidate Response yw

1. On November 21, 2017, President Rodrigo Duterte clarified that he was not going to declare a revolutionary government. He stated that he would continue to work within the framework of the Philippine constitution and existing governmental structures.

2. President Duterte justified this new position by asserting that declaring a revolutionary government was unnecessary and potentially unconstitutional. He explained that the current system already provides ample avenues to address issues and challenges facing the country, such as through executive orders, legislative processes, and other constitutional mechanisms. He emphasized the need to maintain stability and the rule of law, which can be undermined by declaring a revolutionary government.

Figure 48: Response y_w (Instruction Following).

Response yw Grading

[
  {
    "name": "Correctness of the first part of the answer",
    "rating": 2.0,
    "thoughts": "The candidate answer correctly states that President Duterte declared he would not declare a revolutionary government. It aligns with the gold standard answer provided in the rubric."
  },
  {
    "name": "Justification of the new position",
    "rating": 2.0,
    "thoughts": "The candidate answer provides a reasonable explanation for why President Duterte changed his stance, including the assertion that the current system provides sufficient means to address issues and the emphasis on maintaining stability and the rule of law. This closely matches the gold standard answer, which mentions building the country, arresting destabilizers, and maintaining lawfulness."
  }
]

Final score: 2.0

Figure 49: Response y_w Grading (Instruction Following).

Candidate Response yl

1. On November 21, 2017, President Duterte declared that his initial threat to declare a revolutionary government was a bluff and that he had no intention of actually implementing such a move. He stated that the declaration was made under duress due to threats from opposing groups and to protect national interest.

2. President Duterte stated this new position to clarify and de-escalate the tension related to his earlier statements. By clarifying that the threat was not genuine and that he had no plans to implement a revolutionary government, Duterte aimed to reduce political instability and reassure the public and international community that there were no immediate plans for such extreme measures. This action helped to mitigate potential conflicts and maintain stability within the country.

Figure 50: Response y_l (Instruction Following).

Response yl Grading

[
  {
    "name": "Correctness of the first part of the answer",
    "rating": 1.0,
    "thoughts": "The candidate answer correctly states that President Duterte declared he would not declare a revolutionary government. It captures the essence of his clarification without adding incorrect details. However, it does not explicitly mention that the declaration was a bluff or under duress, which is a key part of the gold standard answer."
  },
  {
    "name": "Justification of the new position",
    "rating": 1.0,
    "thoughts": "The candidate answer provides a reasonable explanation for why President Duterte stated his new position, aligning with the gold standard answer. It mentions reducing tension, reassuring the public and international community, and maintaining stability, which are all valid justifications. However, it does not capture the specific points about building the country, arresting destabilizers, and maintaining lawfulness as mentioned in the gold standard."
  }
]

Final score: 1.0

Figure 51: Response y_l Grading (Instruction Following).

Query Synthesis System Prompt (Healthcare QA)

You are a problem proposer. Given a piece of knowledge, create a challenging, well-structured, and unambiguous problem.

Guidelines

* The problem should be assessable.
* The problem should be based on the knowledge and the answer should be derivable from the knowledge. In particular, to create the problem, follow these steps
  1. Extract the list of facts in mentioned in the knowledge.
  2. Create a problem that requires the use of one or more of the extracted facts to answer. Ensure that it is necessary and sufficient to use these facts to derive the correct answer to the problem.
* The problem should ask about a health/medical-related question (pretend that you are a patient and seek health/medical advice; pretend that you are a healthcare professional and ask about professional knowledge, etc.).

Format

* Enclose the problem statement within <problem>...</problem> tags.
  1. In the problem statement, provide the necessary context so that the problem is unambiguous, self-contained, and standalone. Never use: "according to the knowledge", "in the knowledge", "as mentioned", "the knowledge states", "based on the knowledge", etc.
  2. While you should include the necessary context, you should not mention or include any part of the ground truth answer in the problem statement.

* Provide a reference answer within <answer>...</answer> tags. The reference answer should be final and concise, excluding any raw thinking traces.

Figure 52: Query Synthesis Prompt (Healthcare QA).

Query Synthesis System Prompt (Creative Writing)

You are a problem proposer. Given a text sample from the description or main body of a book, create a challenging, well-structured, and unambiguous creative writing question.

Guidelines

* The question should be assessable.
* The question should build upon the given text sample. In particular, to create the question, follow these steps
  1. Extract the key entities, relations, and settings mentioned in the text sample.
  2. Create a creative writing question (write a story, a scene or dialogue in a story, a narrative; expand or continue a story; construct or expand the setting of a story, etc.) that builds upon the text sample and the mentioned entities, relations, and settings.

Format

* Enclose the question statement within <problem>...</problem> tags.
  1. In the question statement, provide the necessary context so that the question is unambiguous, self-contained, and standalone. Never use: "according to the text sample", "in the text sample", "as mentioned", "the text sample states", "based on the text sample", etc.
  2. While you should include the necessary context, you should not mention or include any reference answer in the question statement.
  3. End your question statement with the length requirement "Length: 1000 words.".

* Provide a reference answer within <answer>...</answer> tags. The reference answer should exclude any raw thinking traces.

Figure 53: Query Synthesis Prompt (Creative Writing).

R PROMPTS

We show the prompts in Figure 52 through 64.

Query Synthesis System Prompt (Instruction Following)

You are a problem proposer. Given a piece of knowledge, create a challenging, well-structured, and unambiguous problem.

Guidelines

* The problem should be assessable.

Format

* Enclose the problem statement within <problem>...</problem> tags.
  1. In the problem statement, provide the necessary context so that the problem is unambiguous, self-contained, and standalone. Never use: "according to the knowledge", "in the knowledge", "as mentioned", "the knowledge states", "based on the knowledge", etc.
  2. While you should include the necessary context, you should not mention or include any part of the ground truth answer in the problem statement.

* Provide a reference answer within <answer>...</answer> tags. The reference answer should exclude any raw thinking traces.

Figure 54: Query Synthesis Prompt (Instruction Following).

Query Synthesis User Prompt

<knowledge>
{knowledge}
</knowledge>

Now, please create a challenging, well-structured, and unambiguous problem along with its reference answer, using the provided external knowledge as context.

Figure 55: Query Synthesis User Prompt.

Query Synthesis User Prompt (Creative Writing)

<text_sample>
{knowledge}
</text_sample>

Now, please create a challenging, well-structured, and unambiguous creative writing question along with its reference answer, using the provided text sample as context.

Figure 56: Query Synthesis User Prompt (Creative Writing).

Response Generation System Prompt (Healthcare QA)

You are an assistant that gives health care and medical advice. Answer the given question.

Think step by step, and enclose your final answer in the <answer>...</answer> tags.

Figure 57: Response Generation System Prompt (Healthcare QA).

Response Generation System Prompt (Creative Writing)

You are a helpful assistant. Answer the given question.

Think step by step, and enclose your answer in the <answer>...</answer> tags.

Figure 58: Response Generation System Prompt (Creative Writing).

Response Generation System Prompt (Instruction Following)

You are a helpful assistant. Answer the given question.

Think step by step, and enclose your answer in the <answer>...</answer> tags.

Figure 59: Response Generation System Prompt (Instruction Following).

Response Generation User Prompt

{question}

Figure 60: Response Generation User Prompt.

Rubric Generation System Prompt

Given a piece of knowledge, a problem, its reference answer, and a set of candidate answers, generate a rubric to evaluate the candidate answers. It should consist of 1 to 5 criteria, where each criterion is accompanied by (1) a description, (2) a weight that indicates how important that criterion is to the overall quality, and (3) a gold standard answer if applicable.

Format

Enclose your final rubric in the <rubric>...</rubric> tags.

Within <rubric>...</rubric>, follow the format:

```
<rubric>
{
    "criterion_1": {
        "name": "(Name of criterion_1)",
        "gold": "(The gold standard answer for criterion_1, if applicable.)",
        "description": "(Description of criterion_1. Mention what details to look for to identify good and bad responses for this criterion.)",
        "weight": (Numeric weight for criterion_1)
    },
    ...,
    "criterion_(N)": {
        "name": "(Name of criterion_(N))",
        "gold": "(The gold standard answer for criterion_(N), if applicable.)",
        "description": "(Description of criterion_(N). Mention what details to look for to identify good and bad responses for this criterion.)",
        "weight": (Numeric weight for criterion_(N))
    }
}
</rubric>
```

Reminders

* If all the answers are empty, put 'ERROR' as your response.

* Do not overthink. Finalize it once you come up with a reasonable rubric.

* Each criterion needs to distinguish high-quality answers from low-quality ones meaningfully. Take the following process to identify the suitable criteria:
  1. Group answers (both reference and candidate answers) by quality level. In general, the reference answer should have a high quality compared to the candidate answers, but this is not always true.
  2. Identify factors that separate the answers in high-quality groups from those in low-quality groups.
  3. Select the key factors with the highest discriminative power as the criteria to include in the rubric.

* Each criterion needs to be atomic and focus on a single aspect of quality. Different criteria should not overlap with each other.

* Each criterion needs to be specific to the problem as much as possible. If possible, avoid general criteria that are applicable to any problem (i.e., avoid "factuality", "clarity", "conciseness", etc.). If it is necessary to include general criteria, change them to ones that are specific to the problem (e.g, instead of "factuality", put "factuality of XXX").

* For each criterion, whenever applicable, extract or infer the ground truth or gold standard answer from the provided knowledge, and put it in the "gold" field.
  1. The ground truth or gold standard answer needs to be as concise as possible, ideally a single sentence or phrase (e.g., The criterion is about property "YYY" of "XXX". The ground truth answer is "ZZZ". Put "ZZZ" in the "gold" field.).
  2. If not applicable, put "Not applicable" in "gold".

* For each criterion, its description in the "description" field needs to be as specific to the problem as possible.
  1. Do not just give a general description for the criterion. Elaborate on it with key details, key facts, key phrases, keywords, and examples (e.g., The criterion is about understanding of "XXX". To describe it, instead of saying "The level of understanding of XXX", say "The level of understanding of XXX. Correct understanding should include YYY and ZZZ. ...").
  2. Whenever applicable, connect your description to the gold standard answer that you provide in the "gold" field.

* A small number of high-quality criteria are better than a large number of low-quality criteria. 2 specific, highly distinguishing criteria >> 5 general criteria that rate all answers equally.

Figure 61: Rubric Generation System Prompt.

Rubric Generation User Prompt

<knowledge>
{knowledge}
</knowledge>

<problem>
{question}
</problem>

<reference_answer>
{ref_answer}
</reference_answer>

<candidate_answers>
{answers}
</candidate_answers>

Now generate the rubric to evaluate the candidate answers.

Figure 62: Rubric Generation User Prompt.

Response Grading System Prompt

Given a problem, an evaluation rubric, and a candidate answer, evaluate the candidate answer by giving the answer a rating for each criterion in the rubric.

Format

Put your rubric-based evaluation in <evaluation>...</evaluation> tags.

```
<evaluation>
{
   "criterion_1": {
      "name": "(Name of criterion_1)",
      "thoughts": "(Your thoughts on how well the answer does for criteron_1. If there is a gold standard answer in the "gold" field, does the answer match the gold standard?)",
      "rating": (Your numeric rating between 0 and 2 according to the description and gold standard answer of criteron_1)
   },
   ...,
   "criterion_(N)": {
      "name": "(Name of criterion_(N))",
      "thoughts": "(Your thoughts on how well the answer does for criteron_(N). If there is a gold standard answer in the "gold" field, does the answer match the gold standard?)",
      "rating": (Your numeric rating between 0 and 2 according to the description and gold standard answer of this criteron_(N))
   }
}
</evaluation>
```

Reminder

* The rating needs to be either 0 (Bad), 1 (Medium), or 2 (Good). Higher ratings should indicate better qualities.

* For each criterion that has a gold standard answer in the "gold" field, give a rating of 2 only when the response FULLY matches the gold standard. Give a rating of 1 when the response partially matches the gold standard. Give a rating of 0 when the response does not match any part of the gold standard.

* If the candidate answer is empty, put nothing in the <evaluation>...</evaluation> tags.

Figure 63: Response Grading System Prompt.

Response Grading User Prompt

<problem>
{question}
</problem>

<rubric>
{rubric}
</rubric>

<candidate_answer>
{answer}
</candidate_answer>

Now evaluate the Candidate Answer.

Figure 64: Response Grading User Prompt.