# CROSSTHINK: Scaling Self-Learning beyond Math Reasoning

Anonymous ACL submission

## Abstract

Prior work has successfully applied Reinforcement Learning (RL) to mathematical reasoning—where rules and correctness are well-defined. Yet, generalizing these methods to broader reasoning domains remains challenging due to limited data and the lack of verifiable rewards for unstructured domains. In this work, we propose CROSSTHINK, a framework that systematically incorporates multi-domain corpora into RL training to improve generalization across diverse reasoning tasks. CROSSTHINK addresses key challenges by (1) combining data from varied sources; (2) applying structured templates to control answer-space complexity; (3) filtering for verifiable answers; and (4) optimizing data blending strategies to utilize multi-source data effectively. This enables scalable and verifiable reward modeling beyond math and demonstrates improved accuracies on both math (MATH-500: +30.1%, AMC23: +27.5%) and non-math reasoning benchmarks (MMLU-PRO: +12.8%, GPQA-DIAMOND: +11.3%, AGIEVAL: +15.1%, SUPERGPQA: +3.8%). Moreover, CROSSTHINK exhibits significantly improved response efficiency—using 28% fewer tokens for correct answers—highlighting more focused and effective reasoning. Through CROSSTHINK, we demonstrate that integrating multi-domain, multi-format data in RL leads to more accurate, efficient, and generalizable LLMs.

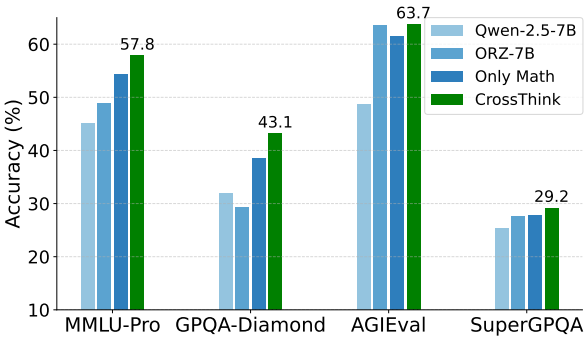


Figure 1: Employing self-learning with multi-domain data, CROSSTHINK outperforms baseline models, including domain-specific training (Only Math) and Open-Reasoner-Zero (ORZ-7B), achieving consistent gains across all reasoning tasks.

## 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable reasoning abilities across a wide range of tasks, with Reinforcement Learning (RL) playing a key role in refining their deep thinking abilities (Hu et al., 2025; Aggarwal and Welleck, 2025; Luo et al., 2025; DeepSeek-AI, 2025; Qin et al., 2024; Huang et al., 2025; Team, 2025c). Recent advances in RL have been particularly successful in mathematical reasoning and coding, where well-defined rules and verifiable correctness enable effective reward modeling. Yet, extending these techniques to broader reasoning domains poses significant challenges, such as—limited training data for RL due to the difficulty of defining verifiable rewards, and ensuring generalization across diverse tasks.

Recent works (Hu et al., 2025; Luo et al., 2025; Cui et al., 2025) have shown ways to diversify RL training corpora by collecting datasets from multiple sources. However, they do not evaluate the relative importance of each source for reasoning or explore optimal data-blending strategies to maximize performance. Furthermore, prior research has largely focused on math reasoning, overlooking the role of non-math reasoning domains in RL training for generalization in out-of-distribution domains. Reasoning process varies across domains and question types. For instance, math problem-solving follows a rule-based, structured, and symbolic approach (Dehaene, 2011), whereas reasoning in fields such as law, physics, social sciences, and history often relies on narrative structures, contextual knowledge, and heuristic strategies. Moreover, different question formats require distinct cognitive approaches—open-ended questions demand

the generation of novel responses, while multiple-choice questions (MCQ) can often be solved by evaluating the given options and selecting the most appropriate answer. Incorporating a diverse range of reasoning domains and question types into RL-based self-learning can enhance the broad reasoning ability of LLMs by exposing them to varied cognitive strategies and knowledge structures.

In this work, we propose CROSSTHINK, a systematic way to incorporate multi-domain corpora for RL training that results in better generalization across a wide variety of tasks. As outlined in Figure 2, CROSSTHINK comprises of phases that—(a) curate data from diverse sources, including synthetic data from web texts and open-source question-answer (QA) pairs on STEM, humanities, law, and social sciences (b) apply templates (MCQ/Open-Ended) to limit the answer-space (d) prepare blends to combine multi-domain data efficiently and (e) employ self-learning with RL to refine reasoning capability in diverse domains.

We evaluate CROSSTHINK along three axes: (1) the effectiveness of data blending strategies in self-learning (2) whether the blending impact amplifies by training with more complex data samples (3) the influence of question-answer templates on downstream accuracies. CROSSTHINK demonstrates that the integration of multi-domain data with varied question formats in RL boosts LLM's reasoning across diverse domains (Figure 1). Notably, models trained with CROSSTHINK not only achieve higher accuracy but also exhibit dynamic response strategies—generating concise answers for general-purpose questions and more detailed responses for math problems—thereby reducing inference cost while preserving task-specific rigor. In addition, our approach addresses the challenge of designing scalable verifiable reward for non-deterministic domains by employing different templates on the curated data to limit the nuances in the answer space diversity. Furthermore, we explore a simple yet effective filtering approach to rank general purpose reasoning (GPR) data based on complexity and show that training with harder samples further amplifies the impact of RL across all domains.

In summary, our contributions are as follows:

- We introduce CROSSTHINK, a novel framework for incorporating multi-domain corpora into RL training, enhancing generalization across diverse reasoning tasks with substantial gains on math (MATH-500: **+30.1%**, AMC23: **+27.5%**) and non-math (MMLU-PRO: **+12.8%**, GPQA-DIAMOND: **+11.3%**, AGIEVAL: **+15.1%**, and SUPERGPQA: **+3.8%**) benchmarks.
- We demonstrate that applying question/answer templates to constrain output diversity leads to more stable reward modeling. Specifically, using a unified open-ended question format improves performance by **1.21%** over mixed-format questions, while short-form answer templates outperform long-form ones by **1.20%**.
- We show that math-only training is insufficient—blending multi-domain data in RL boosts average reasoning accuracy by **1.61%** over math-only data and improves response efficiency by reducing token usage by **28%**.
- We propose a simple yet effective model-driven filtering technique that selects harder samples by removing data solvable by smaller models. This leads to an additional **2.15%** average accuracy gain for `Qwen-2.5-32B`, highlighting the scalability of our approach to larger models.
- We will release **287.4K** high-quality multi-domain data curated for verifiable reward modeling to support future research.

Applying CROSSTHINK on different blends yields substantial improvement over base model (+8.55%-13.36% on average) across seven diverse GPR and math benchmarks. The most effective blend—2:1 ratio of GPR to math data—achieves the highest average accuracy with a 13.36% gain over baseline (Figure 1). Overall, these findings illustrate that thoughtful choices in data blending, scaling, formatting, and filtering are critical to the success of RL with language models. We hope that CROSSTHINK serves as a practical and extensible framework for leveraging multi-domain data to train more capable, reliable, and generalizable models under the RL paradigm.

## 2 CROSSTHINK: Scaling Self-Learning Beyond Math

While mathematical reasoning benefits from clean, verifiable datasets, extending RL to general-purpose reasoning is challenging due to the lack of structured, high-quality supervision. To address this, we leverage web documents and open-source QA benchmarks to collect general-purpose reasoning (GPR) data. However, combining structured and unstructured domains introduces noise and ambiguity—particularly in open-ended formats—making

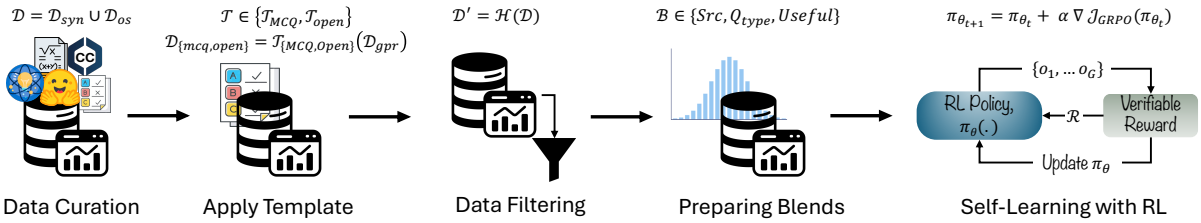


Figure 2: **CROSSTHINK.** We (a) curate QA pairs from from CommonCrawl and open-source datasets, categorized into general-purpose reasoning ($\mathcal{D}_{gpr}$) and mathematical reasoning ($\mathcal{D}_{mr}$); (b) apply structured templates to convert data into MCQ and open-ended formats, promoting diverse reasoning trajectories; (c) filter out unverifiable or ill-formatted responses; (d) deploy RL using Group Relative Policy Optimization (GRPO). The final reward is used to update the policy, iteratively improving the model's reasoning capabilities across diverse domains.

it difficult to apply rule-based reward reliably. To mitigate this, we apply task-specific templates to unify formats, limiting answer space variability and enabling effective verifiable reward signals. Next, we apply a lightweight data filtering to discard unverifiable examples for stable and interpretable RL training. Finally, we explore optimal data blending strategies to investigate how the inclusion of general-purpose reasoning data complements mathematical reasoning, ultimately leading to broader and more adaptive generalization in LLMs.

**Data Curation.** As shown in Table 1, we start with curating datasets from multiple sources to ensure diversity in the training data. Our training data $\mathcal{D}$ comprises of:

$$\mathcal{D} = \mathcal{D}_{syn} \cup \mathcal{D}_{os}$$

Here, $\mathcal{D}_{syn} \rightarrow$ synthetically generated from CommonCrawl (CC) (Gao et al., 2020) and $\mathcal{D}_{os} \rightarrow$ open-source QA datasets. Each source of data further consists of QA pairs related to GPR and math:

$$\mathcal{D}_{\text{syn}} \rightarrow \mathcal{D}_{\text{syn\_gpr}} \cup \mathcal{D}_{\text{syn\_mr}}$$
$$\mathcal{D}_{\text{os}} \rightarrow \mathcal{D}_{\text{os\_gpr}} \cup \mathcal{D}_{\text{os\_mr}}$$

- **General Purpose Reasoning**, $\mathcal{D}_{gpr}$: We collect open source QA datasets ($\mathcal{D}_{os\_gpr}$)—Natural Reasoning (Yuan et al., 2025) and MMLU [Train] (Hendrycks et al., 2021a) that span domains including STEM, Economics, Social Sciences, and more. To enhance diversity, we further synthesize QA pairs from CC documents called CROSSTHINK-QA ($\mathcal{D}_{syn\_gpr}$).

$$\mathcal{D}_{gpr} \rightarrow \mathcal{D}_{syn\_gpr} \cup \mathcal{D}_{os\_gpr}$$

- **Mathematical Reasoning**, $\mathcal{D}_{mr}$: We combine open-source math datasets ($\mathcal{D}_{os\_mr}$): MATH (Hendrycks et al., 2021b) and Numina-Math (Beeching et al., 2024). We generate additional math problems defined as CROSSTHINK-MATH ($\mathcal{D}_{syn\_mr}$) to augment reasoning diversity.

$$\mathcal{D}_{mr} \rightarrow \mathcal{D}_{syn\_mr} \cup \mathcal{D}_{os\_mr}$$

| Data Source | Category | Type | Samples |
|---|---|---|---|
| MMLU [Train] | GPR | MCQ | 99,842 |
| CROSSTHINK-QA* | GPR | MCQ | 192,930 |
| NATURAL REASONING | GPR | OE | 100,000 |
| NuminaMath | MR | OE | 87,350 |
| CROSSTHINK-MATH* | MR | OE | 100,000 |
| Math | MR | OE | 8523 |
| **Total** | | | **588,645** |

Table 1: Training data distribution by source and type. OE=Open-Ended; GPR=General-Purpose Reasoning; MR=Math Reasoning. **We refer to Appendix B for generation details.*

**Applying Templates for Answer Space and Reasoning Diversity.** General purpose reasoning benchmarks are often divided into two categories: (a) Multiple Choice Questions (Hendrycks et al., 2021a; Wang et al., 2024) and (b) Open-Ended Questions (Zhong et al., 2023). Prior works overlooked these variations in the answer space for consistent reward design for tasks which are predominantly math (Hu et al., 2025; Aggarwal and Welleck, 2025; Luo et al., 2025). We hypothesize that each question type elicits different thinking patterns, leading to diverse reasoning trajectories in the model. Therefore, we synthesize $\mathcal{D}_{gpr}$ using two templates: $\mathcal{T}_{MCQ}$ - Multiple Choice Questions (MCQ), and $\mathcal{T}_{Open}$ - Open-Ended questions. We convert the MCQ datasets (MMLU) to open-ended by removing the options from the questions.

$$\mathcal{D}_{mcq} = \mathcal{T}_{MCQ}(\mathcal{D}_{gpr}), \quad \mathcal{D}_{open} = \mathcal{T}_{Open}(\mathcal{D}_{gpr})$$

Additionally, some MCQ questions are incomplete without options (e.g., *Which of the following*

*ways we can file taxes?*). We discard them to avoid confusion during answer generation. Finally,

$$\mathcal{D}_{gpr} = \mathcal{D}_{mcq} \cup \mathcal{D}_{open}$$

**Data Filtering and Formatting.** To obtain high-quality data, we apply a series of filtering and formatting steps, $\mathcal{H}$, to remove samples that are infeasible to evaluate with rule-based reward. Specifically, for $\mathcal{D}_{mcq}$, we check whether the correct answer appears within the question text itself. Given a question-answer pair $(q, a^*)$ with answer choices $\{a_1, a_2, \ldots, a_n\}$, we discard a sample if $a^* \notin \{a_1, a_2, \ldots, a_n\}$. For $\mathcal{D}_{open}$, we discard samples that are challenging to evaluate with a rule-based reward function. Formally, we retain samples where $|w(a^*)| \leq 10$; $w(a^*)$ represents the number of words in the answer $a^*$.

Lastly, for $\mathcal{D}_{mr}$, we remove entries that lack an associated answer, ensuring that all retained questions $q$ have a valid response $a^*$, i.e., we discard samples where $a^* = \emptyset$.

$$\mathcal{D}' = \mathcal{H}(\mathcal{D}) = \{(q, a^*, \{a_1, \ldots, a_n\}) \in \mathcal{D}$$

**Data Blending.** We study the impact of data diversity in three paradigms:

- **Data Source:** We observe the effect of data sources—$\mathcal{D}_{mr}$ and $\mathcal{D}_{gpr}$—by tuning their relative weights in the RL training data.
- **Question Types:** We investigate the impact of question types in downstream tasks.
- **Data Usefulness:** To analyze the contribution of each data source, we run RL using individual data alone and then evaluate them across diverse downstream tasks. Based on their performances, we create a new blend.

Based on these categories, we construct six blends, summarized in Table 10, with their corresponding weight distributions detailed in Table 11.

**Reinforcement Learning with GRPO.** We begin with a pretrained large language model (LLM) $\mathcal{M}$ and a training blend $\mathcal{B}$, where each sample contains only the input prompt and the final answer which is verifiable. We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024). More details can be found in Appendix A.

**Rule Based Reward Modeling.** To guide the RL training, we employ a rule-based reward designed for verifiable evaluation. Similar to (DeepSeek-AI, 2025), we define the total reward function $\mathcal{R} = \mathcal{R}_{\text{acc}} \wedge \mathcal{R}_{\text{format}}$ as the combination of an accuracy reward $\mathcal{R}_{\text{acc}}$ and a format reward $\mathcal{R}_{\text{format}}$. This implies that the output will get reward only when both the answer and the format are correct. Each reward is further detailed in Appendix A

## 3 Experimental Setup

**Training Details.** We adopt `Qwen2.5-7B` and `Qwen2.5-32B` (Team, 2024a) as $\mathcal{M}$, which demonstrate strong generalization capabilities across various reasoning tasks. We directly apply GRPO on $\mathcal{M}$ using the `veRL` framework[1]. We train $\mathcal{M}$ with key settings including a constant learning rate of 1e-6, a batch size and PPO mini batch size of 128 and a maximum context length of 5000 tokens. Each generation step contains 128 unique prompts sampled from the dataset, and performing 8 rollouts with temperature and `top-p` both set to 1.0. We set KL coefficient to 0.001 in all experiments. We conduct training on 4 8 x NVIDIA-H100-80GB nodes, and each training takes approximately 48 GPUs hours.

**Evaluation Metrics.** We evaluate reasoning performance on diverse math and general-purpose benchmarks: MATH-500 (Hendrycks et al., 2021b), AMC23, test set of MMLU (Hendrycks et al., 2021a), MMLU-PRO (Wang et al., 2024), AGIEVAL (Zhong et al., 2023), GPQA-DIAMOND (Rein et al., 2024) and SUPERGPQA (Team et al., 2025). Notably, SUPERGPQA is a recent and rigorous benchmark designed to test the generalizability of LLMs across 285 graduate-level disciplines. Unlike existing benchmarks that concentrate on well-represented domains (e.g., math, law, physics), SUPERGPQA captures long-tail knowledge and includes a wide range of real-world professional disciplines, making it a reliable and discriminative frontier for evaluating generalizability in LLMs. We employ `vllm` (Kwon et al., 2023) as the inference backend, with maximum response length of 5k. For each benchmark, we report accuracy averaged over 3 independent inference runs using greedy decoding.

## 4 Experiments and Results

**Analyze the effect of Individual Datasets.** To design an effective multi-source blend, we first assess the impact of each source on self-learning. This helps prioritize useful sources and down-weight less effective ones. We employ RL using

[1] https://github.com/volcengine/verl

| Data Source | MMLU | MMLU-PRO | GPQA-DIAMOND | AGIEVAL | SUPERGPQA | MATH-500 | AMC23 | Avg |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}$ | 74.20 | 45.00 | 31.82 | 48.59 | 25.36 | 48.30 | 40.00 | 44.75 |
| MMLU [Train] | 69.76 | 38.50 | 32.83 | 47.66 | **27.69** | 22.00 | 5.00 | 34.78 |
| CROSSTHINK-QA | 70.45 | **52.41** | 30.81 | 52.10 | 24.57 | 54.20 | 35.00 | 45.65 |
| Natural Reasoning | 68.89 | 31.33 | 33.33 | 46.65 | 22.44 | 68.60 | 42.50 | 44.82 |
| NuminaMath | 72.94 | 52.05 | **33.84** | **54.39** | 26.97 | 76.20 | **55.00** | **53.06** |
| CROSSTHINK-MATH | 53.99 | 28.08 | 18.69 | 45.69 | 16.92 | 77.20 | 50.00 | 41.51 |
| Math | 63.30 | 31.64 | 21.72 | 51.95 | 18.31 | **78.40** | 50.00 | 45.04 |

Table 2: **Results of Self-Learning on Individual Datasets.** Each row shows the downstream evaluation results after self-learning on a single data source. Results highlight the varying strengths of individual datasets across general-purpose and mathematical benchmarks.

$\mathcal{M}$=`Qwen-2.5-7B` on each dataset separately with a fixed training recipe for consistency. Each model is trained for 250 steps and evaluated on the final checkpoint.

As shown in Table 2, different datasets have varying impacts on downstream accuracies across reasoning benchmarks. `NuminaMath` yields the highest overall average, outperforming the baseline ($\mathcal{M}$) by over 8.30%. While particularly strong on math tasks like MATH-500 and AMC23, it also generalizes well to broader reasoning benchmarks. CROSSTHINK-QA demonstrates a ∼1.0% improvement over baseline with stronger accuracy in MMLU-PRO, AGIEVAL and MATH-500 tasks, suggesting that synthetically generated instruction-style data can generalize well when aligned with benchmark distributions. `Natural Reasoning`, despite modest scores on language-rich benchmarks, delivers a strong average, driven by high scores in MATH-500 and AMC23. This indicates that reasoning-focused datasets, even if less formatted, can contribute meaningfully in math-adjacent tasks. In contrast, CROSSTHINK-MATH performs well on math but generalizes poorly to other domains. Finally, MMLU [Train] underperforms across most tasks, specifically in math domains, suggesting that self-learning with raw MMLU [Train] data alone is insufficient for generalization. However, it excels on SUPERGPQA, which spans cross-disciplinary reasoning, highlighting its potential in capturing broad conceptual knowledge and supporting transfer to long-tail domains—making it a valuable component when targeting general-purpose reasoning benchmarks. While preparing $\mathcal{B}_{score}$, we weight datasets based on their average accuracy—prioritizing sources like CROSSTHINK-QA and `NuminaMath`, while downweighting less effective ones like MMLU [Train].

**Analysis across Blends.** To show the distinction between natural distribution and selective weighting of domains, we prepare $\mathcal{B}_{nd}$, which samples data in proportion to each dataset's original size. Next, to analyze the impact of within-domain vs. cross-domain training, we introduce a `Single Source` category with two domain-specific blends: $\mathcal{B}_{only\_mr}$ and $\mathcal{B}_{only\_gpr}$, using only $\mathcal{D}_{mr}$ and $\mathcal{D}_{gpr}$ respectively. We further compare our approach with a recent math-centric self-learning approach, OPEN-REASONER-ZERO (ORZ) (Hu et al., 2025)—which achieved strong math accuracy using combination of math data. For fair comparison, we evaluate ORZ-7B using our eval setup.

As shown in Table 3, each blend outperforms $\mathcal{M}$ by a significant margin. $\mathcal{B}_{nd}$ yields a 13% average improvement over $\mathcal{M}$, suggesting that simple data diversity—even without rebalancing—can be beneficial. $\mathcal{B}_{gpr\uparrow}$ achieves the highest overall average, with the strongest results across most benchmarks (e.g., MMLU-PRO: +12.82%, AGIEVAL: +15.12%). Notably, it outperforms ORZ by ∼5% on average. While $\mathcal{B}_{only\_mr}$ performs slightly better on math, it lags ∼3–4% behind $\mathcal{B}_{gpr\uparrow}$ on non-math reasoning tasks such as AGIEVAL, SUPERGPQA, and MMLU-PRO. The trend also holds for ORZ. Our analysis with sub-category accuracies in Appendix F reveals that $\mathcal{B}_{gpr\uparrow}$ shows large relative gains in non-math categories while gains in math subcategories are either negligible or even favor $\mathcal{B}_{gpr\uparrow}$ in some tasks. This highlights that multi-domain data offers strong cross-domain transfer with minimal compromise on math accuracy, making it more versatile.

Both $\mathcal{B}_{mcq\uparrow}$ and $\mathcal{B}_{open\uparrow}$ show consistent gains, with the latter achieving a slight edge (+0.6% on average) with stronger results on math tasks. Since math problems are inherently open-ended in structure, highlighting more open-ended domains aligns with the format and reasoning demands of math tasks—leading to better generalize to both general purpose reasoning (GPR) and math tasks. Despite outperforming $\mathcal{M}$, $\mathcal{B}_{score}$ is overall worse than

| Model | Category | Blend | MMLU | MMLU-PRO | GPQA-DIAMOND | AGIEVAL | SUPERGPQA | MATH-500 | AMC23 | Avg |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{M}$ | | | 74.20 | 45.00 | 31.82 | 48.59 | 25.36 | 48.30 | 40.00 | 44.75 |
| ORZ | | | 73.20 | 48.90 | 29.30 | 63.49 | 27.60 | 81.40 | 62.50 | 55.20 |
| CROSSTHINK | | $\mathcal{B}_{nd}$ | 73.18 | 54.81 | 38.07 | 59.99 | 26.54 | 77.00 | 60.00 | 55.66 |
| | Data Source | $\mathcal{B}_{mr\uparrow}$ | 74.85 | 55.51 | 40.10 | 61.47 | 26.81 | 77.80 | 67.50 | 57.72 |
| | | $\mathcal{B}_{gpr\uparrow}$ | **74.94** | **57.82** | 38.58 | **63.71** | **29.16** | 77.60 | 65.00 | **58.12** |
| | Question Types | $\mathcal{B}_{mcq\uparrow}$ | 74.26 | 55.77 | 39.59 | 62.54 | 28.05 | 78.00 | 60.00 | 56.89 |
| | | $\mathcal{B}_{open\uparrow}$ | 74.46 | 55.82 | **43.15** | 61.28 | 26.82 | 78.40 | 62.50 | 57.49 |
| | Data Usefulness | $\mathcal{B}_{score}$ | 74.70 | 56.16 | 40.10 | 59.80 | 27.37 | 78.00 | 62.50 | 56.95 |
| | Single Source | $\mathcal{B}_{only\_mr}$ | 74.24 | 54.26 | 38.58 | 61.39 | 27.69 | **78.60** | **70.00** | 57.82 |
| | | $\mathcal{B}_{only\_gpr}$ | 72.77 | 52.06 | 37.06 | 56.56 | 27.44 | 72.20 | 55.00 | 53.30 |

Table 3: **Results of CROSSTHINK-7B across Blends.** $\mathcal{B}_{gpr\uparrow}$ achieves the highest overall average accuracy, outperforming domain-specific and naturally sampled blends—underscoring the benefit of self-learning with diverse reasoning data.

$\mathcal{B}_{mr\uparrow}$ or $\mathcal{B}_{only\_mr}$. This gap arises because $\mathcal{B}_{score}$ assigns weights based on average scores, without accounting for task-specific strengths. For example, `Math` and CROSSTHINK-MATH are overrepresented due to math performance, while datasets like `MMLU` or `Natural Reasoning`, which excel in general reasoning, are underweighted. In contrast, domain-aware blends selectively prioritize datasets based on their utility within specific domains, leading to more effective coverage and stronger scores across both math and GPR tasks.

In *Single Source* vs. multi-domain analysis, $\mathcal{B}_{only\_mr}$ achieves the highest average math score,ranking as the second-best blend overall in terms of average accuracy. In contrast, while $\mathcal{B}_{only\_gpr}$ outperforms $\mathcal{M}$, it underperforms in math tasks and trails 4.2% on average across non-math reasoning tasks, despite being tailored for GPR. This counterintuitive finding suggests that to obtain maximum gain in GPR tasks we need to include math problems in the training. As discussed earlier, $\mathcal{B}_{gpr\uparrow}$ gets the best average reasoning accuracy which consists of both math and GPR domains. This confirms that math data alone is transferable to structured reasoning tasks, whereas GPR data is less effective when isolated.

## 5 Ablations

**CROSSTHINK is token efficient in responses.** To further understand the influence of multi-domain data in response generation, we compare the average token lengths of correct and incorrect responses between models trained on two blends: $\mathcal{B}_{gpr\uparrow}$ and $\mathcal{B}_{only\_mr}$. As shown in Figure 3, on general-purpose reasoning (GPR) benchmarks, $\mathcal{B}_{gpr\uparrow}$ consistently outperforms $\mathcal{B}_{only\_mr}$ and ORZ (Hu et al., 2025), not only in accuracy (as shown in Table 3) but also in response efficiency—producing correct answers with significantly fewer tokens[2]. For instance, on MMLU, the average token count for correct responses is 229 for $\mathcal{B}_{gpr\uparrow}$, compared to 351 for $\mathcal{B}_{only\_mr}$. This demonstrates that exposure to multi-domain data enables the model to internalize a more efficient reasoning strategy, leading to both improved performance and reduced inference cost.

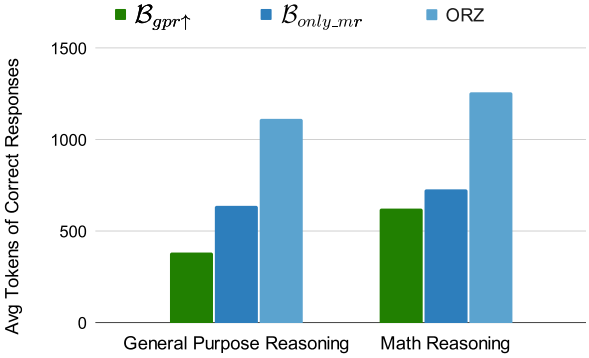


Figure 3: Token efficiency comparison of models trained on $\mathcal{B}_{gpr\uparrow}$ (multi-domain blend) and two single domain blends ($\mathcal{B}_{only\_mr}$ and ORZ).

In contrast, on math-specific benchmarks, $\mathcal{B}_{only\_mr}$ and ORZ perform slightly better in accuracy, as expected due to domain alignment. Interestingly, correct responses are generally longer than GPR tasks as solving math problems inherently requires detailed, multi-step derivations, hypothesis exploration, verification and refinement. Despite this, the $\mathcal{B}_{gpr\uparrow}$ shows its adaptability by generating longer responses for math tasks and shorter ones for GPR tasks—indicating a dynamic response strategy learned through multi-domain training. As shown in Table 12, $\mathcal{B}_{gpr\uparrow}$ increases its average tokens by 62% when generating responses for math tasks (Mean Tokens=622) as opposed to GPR tasks

[2]Detailed categorization per task is shown in Appendix E.

(Mean Tokens=385). Whereas, $\mathcal{B}_{only\_mr}$ increases by 14% (Mean Tokens=731 for math and Mean Tokens=639 for GPR tasks) showing a much smaller dynamic range. This trend is also mirrored in ORZ which shows an even smaller increase (12%) in average token length across domains.

This adaptive behavior highlights a key strength of multi-domain training: it equips the model with the flexibility to tailor its response style to the nature of the task. By learning from a diverse range of domains, $\mathcal{B}_{gpr\uparrow}$ learns to reason efficiently—across all tasks, $\mathcal{B}_{gpr\uparrow}$ uses on average 28% fewer tokens for correct responses than $\mathcal{B}_{only\_mr}$—producing compact yet accurate answers where appropriate, and detailed ones when necessary.

**Data Format Study: Question and Answer Templates.** To examine the training data formatting effect on model performance, we conduct two controlled studies focused on question and answer template design. In Table 3, we observe that $\mathcal{B}_{open\uparrow}$ outperforms $\mathcal{B}_{mcq\uparrow}$, suggesting that models trained on more open-ended data generalize better across benchmarks. This motivated us to investigate whether converting all questions into a unified open-ended format leads to better performance. In ***Question Template Study***, we use the natural distribution blend ($\mathcal{B}_{nd}$) and only perturb the question template. To generate the open-ended variant, we remove the answer options from MCQs, prompting the model to produce an answer without selecting from predefined choices.

| Question Type | GPR Avg | Math Avg | Total Avg |
|---|---|---|---|
| MCQ +OPEN-ENDED | 50.52 | 68.50 | 55.66 |
| OPEN-ENDED | **51.30** | **70.80** | **56.87** |

Table 4: **Impact of Question Format.** Converting all questions to open-ended format improves accuracy across benchmarks, reducing reliance on option guessing and encouraging deeper reasoning (Appendix C).

From Table 4, the open-ended setting surpasses the mixed-format one on nearly all tasks, achieving 1.21% higher average score. It yields notable gains on reasoning-intensive and MCQ-formatted benchmarks such as MMLU, SUPERGPQA, and GPQA-DIAMOND. This result may be attributed to the inherent structure of MCQ questions, where random guessing can yield an accuracy of approximately 25% in MMLU and GPQA-DIAMOND where we have four options. In contrast, open-ended questions eliminate this guessing advantage, compelling the model to rely heavily on reasoning to arrive at a correct answer. By reducing the likelihood of reward hacking through random option selection, the open-ended format encourages more robust reasoning and leads to improved generalization.

In ***Answer Template Study***, we investigate how the format of MCQ-style output labels influences training. We compare two answer templates: *Long* - the model is trained to generate both the option label and its corresponding description (e.g., `(A) Sky is blue`), and *Short* - the model is trained to output only the option label (e.g., `A`). Here, we use the $\mathcal{B}_{only\_gpr}$ blend, which primarily consists of MCQ datasets (Table 1), making it ideal for analyzing the effects of answer formatting in this setting.

| Answer Type | GPR Avg | Math Avg | Total Avg |
|---|---|---|---|
| Long | 49.18 | **63.60** | 53.30 |
| Short | **50.95** | 63.35 | **54.50** |

Table 5: **Impact of Answer Format.** Using short-form answers improves accuracy by reducing output ambiguity and avoiding penalization from rigid reward functions in rule-based training (Appendix C).

As shown in Table 5, the short answer template outperforms the long-form variant, with a 1.20% gain in average accuracy. The trend holds for both GPR and math benchmarks. These results suggest that reducing the complexity of the output space helps minimize ambiguity and allows the model to better align its predictions with the structure of the question. Furthermore, when training with long-form answers using a rule-based reward (e.g., exact string matching), the model is often penalized for minor deviations in phrasing, even when the correct option is selected. This introduces noisy supervision and may hinder learning. While this issue could be mitigated by designing a more flexible reward function (e.g., LLM-as-a-Judge), we aim to keep our approach simple and interpretable. As such, we adopt a naive rule-based reward for clarity and reproducibility, and leave more sophisticated reward designs for future investigation.

**Difficulty Filtering.** High-quality data is a key factor in self-learning to ensure efficient and stable learning. Recent works (Hu et al., 2025; Luo et al., 2025; Cui et al., 2025; Zeng et al., 2025; Fatemi et al., 2025) investigate data selection based on question complexity, showing that training on harder questions improves downstream accuracy. However, their approach relies on datasets with pre-

defined difficulty scores. In this work, we explore a simple approach to estimate question difficulty for GPR datasets that do not come with explicit difficulty labels. Specifically, we label questions as 'difficult' if they are answered incorrectly by a smaller model (`Qwen-2.5-7B`) in a zero-shot setting and filter out the 'easy' questions. The intuition is that questions easily answered by a base model are likely to be knowledge-based or shallow in reasoning depth, whereas those it fails on are likely to require deeper reasoning or broader generalization. We construct two versions of our training dataset $\mathcal{B}_{gpr\uparrow}$—an unfiltered set containing all questions, and a filtered set ($\mathcal{B}_{f(gpr)\uparrow}$) that retains only the difficult samples—and use them to train separate instances of a larger $\mathcal{M}$ = `Qwen-2.5-32B`.

| Model | Blend | GPR Avg | Math Avg | Total Avg |
|---|---|---|---|---|
| Qwen-2.5-32B | | 54.95 | 52.78 | 54.33 |
| CROSSTHINK-32B | $\mathcal{B}_{gpr\uparrow}$ | 62.20 | 74.95 | 65.84 |
| | $\mathcal{B}_{f(gpr)\uparrow}$ | **63.39** | **79.50** | **67.99** |

Table 6: **Difficulty-Based Filtering.** Filtering $\mathcal{B}_{gpr\uparrow}$ to retain only hard examples ($\mathcal{B}_{f(gpr)\uparrow}$) yields consistent gains across all tasks, highlighting the effectiveness of selective training on challenging data (Appendix C).

According to Table 6, this filtering approach results in consistent performance gains across all evaluated benchmarks. While both filtered and unfiltered models outperform $\mathcal{M}$, $\mathcal{B}_{f(gpr)\uparrow}$ achieves the highest accuracy on every task. The gains are especially prominent in complex benchmarks such as MMLU-PRO, GPQA-DIAMOND, AGIEVAL, and AMC23, where $\mathcal{B}_{f(gpr)\uparrow}$ improves by up to 2–8% over $\mathcal{B}_{gpr\uparrow}$. On average, filtering boosts overall accuracy by 2.15%, a notable gain considering that it comes from training on fewer but harder examples. This suggests that selectively training on challenging examples can yield more robust and generalizable models, likely due to stronger gradient signals and a focus on harder-to-learn reasoning patterns.

## 6 Related Work

**Reasoning in LLM.** Large Language Models have achieved strong performance on various NLP tasks, with Chain-of-Thought (CoT) prompting (Wei et al., 2022) enabling multi-step reasoning across domains like math, science, and programming. Long CoT (OpenAI, 2024) further enhances reasoning by introducing behaviors such as reflection, verification, and correction, with strong scaling properties. Models like QwQ (Team, 2024b, 2025c), DeepSeek-R1 (DeepSeek-AI, 2025), Kimi k1.5 (Team, 2025a), and InternThinker (Cai et al., 2024) leverage Long CoT with RL to boost reasoning performance. Smaller models like Open-Reasoner-Zero (Hu et al., 2025), Open-R1 (Face, 2025), O1 (Qin et al., 2024; Huang et al., 2025), s1 (Muennighoff et al., 2025), and LIMO (Ye et al., 2025) also benefit from Long CoT via distillation.

**Data Sampling in RL.** Recent work explores mixing data from multiple sources in RL to improve reasoning diversity and generalization in RL (Hu et al., 2025; Luo et al., 2025; Zeng et al., 2025; Wen et al., 2025). However, they primarily focus on math due to the ease of designing verifiable rewards. Sampling strategies often rely on question complexity or algorithmic verifiability, such as (Xie et al., 2025) uses synthetic puzzles to control difficulty. However, these methods remain limited to structured domains like math or logic. Yeo et al. (2025) reports the best MMLU-PRO scores by blending multi-domain (Yue et al., 2024) data, though majority of it is math-focused, leaving unclear the contribution of non-math data. Despite these efforts, the impact of including non-math domains—like law, social science, or commonsense reasoning—remains underexplored. CROSSTHINK is the first systematic framework to incorporate multi-domain, multi-format data into RL, introducing verifiable rewards for non-deterministic tasks and demonstrating that diverse blends lead to LLMs that reason more broadly, adapt dynamically, and think more efficiently.

## 7 Conclusion

We present CROSSTHINK, a simple and scalable framework for improving the generalization abilities of LLMs through RL with multi-domain corpora. By combining multi-domain data, structured templates, and difficulty-aware filtering, CROSSTHINK enables consistent gains across both general-purpose (+3.8–15.1%) and mathematical (+27.5–30.1%) benchmarks—using 28% fewer tokens for correct responses. Importantly, these benefits persist across model scales and task types, demonstrating that data diversity, not just data volume, is key to broader reasoning capabilities. CROSSTHINK offers a practical recipe for building more generalizable, efficient, and reliable LLMs under the RL paradigm—paving the way for scalable self-learning beyond math.

## 8 Limitations

While CROSSTHINK demonstrates strong improvements in reasoning accuracy, adaptability, and efficiency, there is still room for improvement. As discussed in Section 5, the reward modeling framework used in this work is rule-based and relatively simplistic. Specifically, it relies on exact string matching for correctness and formatting verification, which can be brittle for open-ended responses. For example, if the ground truth is `(A) Sky is blue`, and the model predicts `(A) the sky is generally blue most times`, the answer is semantically correct but still receives a negative reward. This limitation improperly affects general-purpose reasoning tasks with inherently more diverse and less deterministic answer spaces. Future work could incorporate more flexible, semantics-aware reward functions, such as fuzzy matching, entailment scoring, embedding-based similarity metrics, or `llm-as-a-Judge`, to better align reward signals with human judgment. Additionally, we did not perform extensive hyperparameter tuning for RL training. All models were trained using fixed schedules and standard values for learning rate, KL coefficients, and rollout configurations. Moreover, scaling RL training for longer steps and dataset are computationally expensive which constrained us to deploy all runs for a fixed number of steps (650 steps). Recent works (Yu et al., 2025; Aggarwal and Welleck, 2025; Luo et al., 2025) shows improvement over naive GRPO by adjusting clip ratio, dynamic sampling, number of rollouts, context length. While our results are strong under fixed conditions, additional gains may be possible with better-tuned training regimes by exploiting hyperparameters.

## 9 Ethical Considerations

Reinforcement learning strongly incentivizes the reasoning capabilities of LLMs, enabling models to perform better across complex tasks. However, this process can also inadvertently amplify existing biases present in the base model or introduced through reward modeling and data selection. In this work, we primarily focus on scaling and diversifying reasoning ability across domains and do not explicitly address fairness, bias mitigation, or value alignment. Future work should systematically evaluate how reinforcement learning affects the model's behavior across sensitive axes such as gender, race, and geopolitical context—especially in open-ended, non-verifiable tasks.

## References


Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *Preprint*, arXiv:2503.04697.

Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. 2024. Numinamath 7b cot. `https://huggingface.co/AI-MO/NuminaMath-7B-CoT`.

Zheng Cai and 1 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Stanislas Dehaene. 2011. *The number sense: How the mind creates mathematics*. OUP USA.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. 2025. Concise reasoning via reinforcement learning. *Preprint*, arXiv:2504.05185.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. 2025. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey – part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL. Notion Blog.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2025. General-reasoner: Advancing llm reasoning across all domains. https://github.com/TIGER-AI-Lab/General-Reasoner/blob/main/General_Reasoner.pdf.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, and 64 others. 2024. Nemotron-4 340b technical report. *Preprint*, arXiv:2406.11704.

OpenAI. 2024. Learning to reason with llms.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and 1 others. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

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

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

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

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. 2025. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *Preprint*, arXiv:2503.23829.

Kimi Team. 2025a. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, and 77 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Preprint*, arXiv:2502.14739.

Qwen Team. 2024a. Qwen2.5: A party of foundation models.

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

Qwen Team. 2025b. Qwen2.5-vl.

Qwen Team. 2025c. Qwq-32b: Embracing the power of reinforcement learning.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Preprint*, arXiv:2402.10176.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. 2024. Language models learn to mislead humans via rlhf. *Preprint*, arXiv:2409.12822.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *Preprint*, arXiv:2503.10460.

Lilian Weng. 2024. Reward hacking in reinforcement learning. *lilianweng.github.io*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2502.14768.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *Preprint*, arXiv:2311.04850.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long chain-of-thought reasoning in llms. *Preprint*, arXiv:2502.03373.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, and Xian Li. 2025. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *Preprint*, arXiv:2502.13124.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

## A Reinforcement Learning

We utilize Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our RL algorithm. Unlike PPO (Schulman et al., 2017), GRPO does not use a separate critic model and instead estimates the baseline from group scores, improving efficiency and reducing memory. For each question $q$, GRPO samples a group of outputs $o_1, o_2, ..., o_G$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model $\pi_\theta$ by maximizing the following objective:

$$x_{i,t} = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$$

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \; & E\left[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)\right] \\ & \times \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Bigg[\min\Bigg(x_{i,t}\hat{A}_{i,t}, \\ & \text{clip}\Big(x_{i,t}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_{i,t}\Bigg) \\ & -\beta D_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big)\Bigg]\end{aligned}$$

$$\begin{aligned}D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] = & \frac{\pi_{\text{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} \\ & - \log\frac{\pi_{\text{ref}}(o_{i,t}|q, o_{i,<t})}{\pi_\theta(o_{i,t}|q, o_{i,<t})} - 1.\end{aligned}$$

where $\epsilon$ and $\beta$ are hyperparameters, and $\hat{A}_{i,t}$ is the advantage, computed using a group of rewards $\{r_1, r_2, ..., r_G\}$ corresponding to the outputs within each group:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\})}$$

**Defining Rewards.** We combine accuracy reward ($\mathcal{R}_{\text{acc}}$) and format reward ($\mathcal{R}_{\text{format}}$) to estimate the final reward:

**Accuracy Reward:** The accuracy reward evaluates correctness based on whether the model's response $p$ is similar to the ground truth solution $a$ to satisfy the correctness criteria:

$$\mathcal{R}_{\text{acc}}(p, a) = \begin{cases} 1, & \text{if equal}(p, a), \\ 0, & \text{otherwise.} \end{cases}$$

**Format Reward:** The format reward ensures the response $a$ is structured according to predefined tags, where the reasoning will reside in '<think></think>' tokens and the final answer will be shown inside `\boxed{}`:

$$R_{\text{format}}(a) = \begin{cases} 1, & \text{if } F(a), \\ 0, & \text{otherwise.} \end{cases}$$

where $F(a)$ returns `True` if $a$ is correctly formatted and `False` otherwise.

# B Data Synsthesis

To obtain a balanced high-quality multi-domain data, we synthesize a high-quality question answering dataset spanning over math (CROSSTHINK-MATH) and general purpose reasoning (CROSSTHINK-QA) domain using publicly available datasets such as CommonCrawl (Gao et al., 2020). Our dataset is intended to be used by the community to deploy reinforcement learning with LLMs which is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0)[3]. The data may be used to train and evaluate. This dataset contains synthetic data created using `Qwen/Qwen2.5-Math-72B`, `Qwen2.5-72B-Instruct`. If this dataset is used to create, train, fine tune, or otherwise improve an AI model, which is distributed or made available, such AI model may be subject to redistribution and use requirements in the Qwen License Agreement[4].

## B.1 CROSSTHINK-QA

We synthetically generate a large-scale multiple-choice question (MCQ) dataset following two approaches below:

[3] https://creativecommons.org/licenses/by/4.0/legalcode

[4] https://huggingface.co/Qwen/Qwen2.5-Math-72B/blob/main/LICENSE and https://huggingface.co/Qwen/Qwen2.5-72B-Instruct/blob/main/LICENSE

### B.1.1 SDG from Scratch

**Topic, Subtopic, and Difficulty Definition:** We first define a broad set of topics, such as physics, biology, chemistry, and others. For each topic, we use `Nemotron-4-340B-Instruct` (Nvidia et al., 2024) to generate a list of popular subtopics. We also define multiple difficulty levels to ensure diversity and scale of the data.

**Question Generation:** We initially generate few-shot examples that demonstrate various levels of difficulty using `Nemotron-4-340B-Instruct` (Nvidia et al., 2024). We later prompt `Qwen2.5` models (Team, 2024a) along with the few-shot examples to generate a multiple-choice question based on the specified topic, subtopic, and difficulty. Each generated question is checked to ensure it follows the required format.

**Augmentation:** Similarly to the OpenMathInstruct (Toshniwal et al., 2024) pipeline, we augment generated questions by prompting `Qwen2.5` (Team, 2024a) models to create a question similar to or inspired by the original.

**Benchmark Decontamination:** We perform decontamination against test sets of popular MCQ benchmarks such as GPQA (Rein et al., 2024), MMLU (Hendrycks et al., 2021a), and MMLU-PRO (Wang et al., 2024), using the method from (Yang et al., 2023).

**Solution Generation:** We prompt DeepSeek-R1 (DeepSeek-AI, 2025) to generate multiple reasoning traces per question. Since there are no ground-truth answers for the questions generated in earlier stages, we use majority voting over generated solutions to determine the most likely correct answer.

### B.1.2 SDG from Book

We utilized `Qwen2.5-VL-72B-Instruct` (Team, 2025b) for extracting text from textbooks (*OpenStax*[5] and *An Introduction to Formal Logic*[6]), which was then manually checked for transcription accuracy. We employed `Mixtral-8x22B-Instruct-v0.1` (Jiang et al., 2024) and `Qwen2.5-72B-Instruct` (Team, 2024a) to synthesize multiple-choice questions based on

[5] https://openstax.org/

[6] https://forallx.openlogicproject.org/forallxyyc.pdf

sections and key terms extracted from these textbooks. We prompted the models to generate questions with four distinct and plausible options, along with the correct answer and a justification for it. Subsequently, each generated question was evaluated using another model to ensure that every example is self-contained and accurate.

### B.2 CROSSTHINK-MATH

To construct CROSSTHINK-MATH, we adopt an approach similar to the technique in (Ge et al., 2024). Specifically, we use web documents from Common Crawl to build personas, and we use `Qwen2.5-72B-Instruct` (Team, 2024a) model when generating personas. Then, to promote diversity in math questions, we incorporate math skills introduced in (Didolkar et al., 2024) and condition the `Qwen2.5-72B-Instruct` model on both the math skills and the personas. Finally, we use the `Qwen2.5-72B-Math-Instruct` model to generate the solutions. Prompt templates are shown in Figure 4, Figure 5, and Figure 6.

## C Breakdown of Performance

We further provide a breakdown of the results showing impact of data formats and filtering in RL training. Table 7 and Table 8 shows the impact of the question and answer formats across all tasks. In Table 9, we further extend the results for models trained on unfiltered and filtered datasets.

## D Data Proportion across Blends

CROSSTHINK has been trained using the datasets shown in Table 1. To better understand the data composition used in our reinforcement learning experiments, we report the proportion of each dataset in the six blending strategies in Table 10, introduced in Section 2. These proportions reflect how data is distributed across different sources depending on the specific blending paradigm: data source, question type, and data usefulness.

## E Token Efficiency Analysis

**Token Efficiency in Correct Responses.** Understanding not only whether a model answers correctly but also how efficiently it reasons is critical in real-world deployments, especially for reducing inference cost and latency. To this end, we analyze the token lengths of correct responses generated by models trained under different data blending strategies.

Table 12 presents the minimum, maximum, and mean number of tokens used in correct answers across two task types: General Purpose Reasoning (GPR) and Math. We compare three models: (1) $\mathcal{B}_{gpr\uparrow}$ (multi-domain training), (2) $\mathcal{B}_{only\_math}$ (math-only training), and (3) ORZ (a strong math-centric baseline model).

Across GPR tasks, $\mathcal{B}_{gpr\uparrow}$ produces the most concise correct responses, with a mean of 385 tokens—39.6% fewer than $\mathcal{B}_{only\_mr}$ and 65.4% fewer than ORZ. This suggests that training with multi-domain corpora equips the model to reason more efficiently in less structured tasks, avoiding unnecessarily verbose responses.

On math benchmarks, where detailed step-by-step derivations are essential, all models naturally generate longer outputs. However, $\mathcal{B}_{gpr\uparrow}$ still demonstrates adaptability, producing appropriately longer responses compared to GPR, while keeping the output concise relative to $\mathcal{B}_{only\_math}$ and ORZ. This behavior underscores the ability of multi-domain trained models to dynamically adjust their reasoning strategy and verbosity based on task requirements.

Interestingly, ORZ exhibits the longest response lengths across both GPR and math tasks. While this aligns with its design as a reasoning-heavy model, it also reflects less efficiency—potentially generating unnecessarily long chains of thought, particularly in domains outside its training focus.

In summary, the token efficiency analysis reveals that $\mathcal{B}_{gpr\uparrow}$ achieves a favorable trade-off between accuracy and brevity, tailoring its reasoning depth to the complexity of the task. This reinforces the value of diverse, multi-domain training in promoting adaptable and cost-efficient language models.

**Thinking Long vs Thinking Accurate.** Recent studies such as DeepScaler (Luo et al., 2025) have noted that incorrect answers often exhibit longer trajectories, leading to wasted computation and less efficient learning. Echoing this observation, we analyze the average token lengths of correct and incorrect responses for models trained on different blends: $\mathcal{B}_{gpr\uparrow}$, $\mathcal{B}_{only\_mr}$, and ORZ.

As shown in Figure 7, incorrect responses are consistently and substantially longer than correct ones—by 3.6× on average. This pattern holds across both general-purpose and math reasoning tasks, suggesting that verbose reasoning does not guarantee correctness. In fact, longer responses often reflect the model's uncertainty, overthinking,

**Text-to-persona template**

Who is likely to read the text?

{text}

Note: 1. Your response should always start with “Persona:”. 2. The persona should be realistic and detailed, but don’t include specific name of person. 3. Don’t include any preamble or disclaimer, but only provide the persona. 4. Persona should be at most two sentences.

Persona:

Figure 4: Prompt template for text-to-persona generation.

**Persona-to-persona template**

Who is in close relationship with the given persona?

{persona}

Note: 1. Your response should always start with "Related Persona:". 2. The persona should be realistic and detailed, but don’t include specific name of person. 3. Don’t include any preamble or disclaimer, but only provide the persona. 4. Persona should be at most two sentences.

Related Persona:

Figure 5: Prompt template for persona-to-persona generation.

**Persona and skill to math problem template**

Create a math problem related to the following persona and require understanding of the following skills:

Skills: {skills}

Persona: {persona}

Note: 1. The math problem should be challenging and involve given advanced mathematical skills. Only top talents can solve it correctly. 2. You should make full use of the persona description to create the math problem to ensure that the math problem is unique and specific to the persona. 3. Your response should always start with "Math problem:". Your response should not include a solution to the created math problem. 4. Your created math problem should include no more than 2 sub-problems.

Math problem:

Figure 6: Prompt template for persona and skill to problem generation.

| Question Type | MMLU | MMLU-PRO | GPQA-DIAMOND | AGIEVAL | SUPERGPQA | MATH-500 | AMC23 | Avg |
|---|---|---|---|---|---|---|---|---|
| MCQ +OPEN-ENDED | 73.18 | **54.81** | 38.07 | **59.99** | 26.54 | **77.00** | 60.00 | 55.66 |
| OPEN-ENDED | **74.61** | 54.36 | **39.09** | 59.30 | **29.16** | 76.60 | **65.00** | **56.87** |

Table 7: **Impact of Question Format.** Converting all questions to open-ended format improves accuracy across benchmarks, reducing reliance on option guessing and encouraging deeper reasoning.

| Answer Type | MMLU | MMLU-PRO | GPQA-DIAMOND | AGIEVAL | SUPERGPQA | MATH-500 | AMC23 | Avg |
|---|---|---|---|---|---|---|---|---|
| Long | 72.77 | 52.06 | 37.06 | 56.56 | 27.44 | 72.20 | **55.00** | 53.30 |
| Short | **74.22** | **54.56** | **39.59** | **58.01** | **28.39** | **74.20** | 52.50 | **54.50** |

Table 8: **Impact of Answer Format.** Using short-form answers improves accuracy by reducing output ambiguity and avoiding penalization from rigid reward functions in rule-based training.

| Model | Blend | MMLU | MMLU-PRO | GPQA-DIAMOND | AGIEVAL | SUPERGPQA | MATH-500 | AMC23 | Avg |
|---|---|---|---|---|---|---|---|---|---|
| Qwen-2.5-32B | | 83.30 | 55.10 | 40.40 | 62.77 | 33.16 | 60.55 | 45.00 | 54.33 |
| CROSSTHINK-32B | $\mathcal{B}_{gpr\uparrow}$ | 83.57 | 68.83 | 46.70 | 73.90 | 37.99 | 82.40 | 67.50 | 65.84 |
| | $\mathcal{B}_{f(gpr)\uparrow}$ | **83.60** | **69.43** | **49.75** | **75.82** | **38.34** | **84.00** | **75.00** | **67.99** |

Table 9: **Difficulty-Based Filtering.** Filtering $\mathcal{B}_{gpr\uparrow}$ to retain only hard examples ($\mathcal{B}_{f(gpr)\uparrow}$) yields consistent gains across all tasks, highlighting the effectiveness of selective training on challenging data.

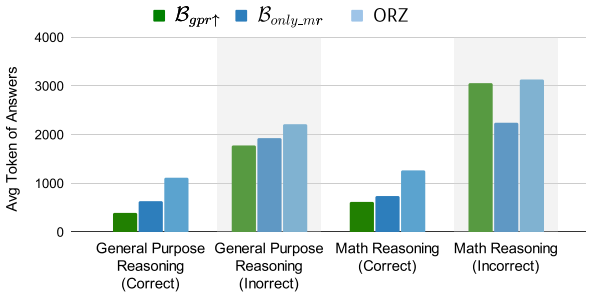


Figure 7: Average token lengths of correct and incorrect responses across general-purpose and math reasoning tasks for models trained on $\mathcal{B}_{gpr\uparrow}$, $\mathcal{B}_{only\_mr}$, and ORZ.

or repetitive CoT traces, rather than productive deduction.

## F Sub-category Accuracy Analysis

To further support our observation that multi-domain training improves general-purpose reasoning while remaining competitive on math tasks, we analyze the number of correct responses across sub-categories in MMLU-PRO and AGIEVAL. Figure 8 and Figure 9 show the count of correct answers produced by $\mathcal{B}_{gpr\uparrow}$ and $\mathcal{B}_{only\_math}$ across their respective sub-domains.

On MMLU-PRO, $\mathcal{B}_{gpr\uparrow}$ consistently outperforms $\mathcal{B}_{only\_math}$ across non-math reasoning categories such as business, law, psychology, chemistry, and economics. Notably, it achieves relative improvements of +20.58% in law and +13.26% in business. Surprisingly, $\mathcal{B}_{gpr\uparrow}$ also performs better in the math category (+7.2%), despite not being trained exclusively on mathematical data. This may be attributed to the nature of MMLU-PRO's math problems, which are college-level and benefit from a combination of symbolic and heuristic reasoning—skills reinforced through exposure to diverse domains.

In contrast, the AGIEVAL benchmark (shown in Figure 9) features Olympiad-level math questions that are more abstract and complex. Here, $\mathcal{B}_{only\_math}$ has a slight edge (+1.8%) in the math category, which aligns with its domain-specific training. However, $\mathcal{B}gpr\uparrow$ demonstrates stronger performance in symbolic and language-heavy domains, showing +13.06% improvement in Law and +9.88% in English. Averaged across all non-math reasoning categories, $\mathcal{B}_{gpr\uparrow}$ achieves a +8.6% relative gain over $\mathcal{B}_{only\_math}$, reinforcing its advantage in general-purpose and real-world reasoning tasks.

A similar trend is observed in the SUPERGPQA sub-category analysis shown in Figure 10. $\mathcal{B}_{gpr\uparrow}$ significantly outperforms $\mathcal{B}_{only\_math}$ across nearly all categories—especially in engineering, agronomy, economics, education, law, and philosophy. The only exception is the "Science" category, which includes math-heavy disciplines like physics, chemistry, and astronomy, where both blends perform comparably. This further highlights that multi-domain training with $\mathcal{B}_{gpr\uparrow}$ enhances reasoning across a broad spectrum of fields, achieving strong generalization even in real-world, professional domains that fall outside traditional math tasks.

## G Extended Related Work

**Self-Learning beyond Math.** High-quality training data are crucial for scalable Reasoner-Zero training. Most of the recent works emphasize mathematical benchmark-centric data (AMC, AIME, Math, Olympiads, and AoPS) for reinforcement learning (Hu et al., 2025; Aggarwal and Welleck, 2025; Trung et al., 2024; Ye et al., 2025; Zeng et al., 2025) as designing verifiable rewards is much easier for math tasks. They exclude problems such as

| Category | Blend Name | Symbol | Blend Description |
|---|---|---|---|
| Data Source | Natural Distribution | $\mathcal{B}_{nd}$ | Ratio of number of samples in a dataset divided by the total number of samples in all the datasets. |
| | More Math | $\mathcal{B}_{mr\uparrow}$ | 2:1 ratio of $\mathcal{D}_{mr}$ and $\mathcal{D}_{gpr}$ |
| | More General Purpose Reasoning | $\mathcal{B}_{gpr\uparrow}$ | 2:1 ratio of $\mathcal{D}_{gpr}$ and $\mathcal{D}_{mr}$ |
| Question Types | More MCQ | $\mathcal{B}_{mcq\uparrow}$ | 2:1 ratio of $\mathcal{D}_{mcq}$ and $\mathcal{D}_{open}$ |
| | More Open-Ended | $\mathcal{B}_{open\uparrow}$ | 2:1 ratio of $\mathcal{D}_{open}$ and $\mathcal{D}_{mcq}$ |
| Data Usefulness | Avg. Score | $\mathcal{B}_{score}$ | Provide weight to each source based on their average benchmark performances |

Table 10: **Overview of Data Blending Strategies.** Blends are categorized by data source, question type, and usefulness—each constructed to assess the impact of domain diversity, format variation, and task relevance on RL-based reasoning.

| Data Name | Type | $\mathcal{B}_{nd}$ | $\mathcal{B}_{mr\uparrow}$ | $\mathcal{B}_{mcq\uparrow}$ | $\mathcal{B}_{open\uparrow}$ | $\mathcal{B}_{gpr\uparrow}$ | $\mathcal{B}_{score}$ | $\mathcal{B}_{only\_math}$ | $\mathcal{B}_{only\_gpr}$ |
|---|---|---|---|---|---|---|---|---|---|
| MMLU | MCQ | 0.1696 | 0.0864 | 0.2251 | 0.1159 | 0.1678 | 0.1296 | | 0.2542 |
| CrossThink-QA | MCQ | 0.3277 | 0.1670 | 0.4349 | 0.2241 | 0.3242 | 0.1731 | | 0.4912 |
| Natural Reasoning | OPEN-ENDED | 0.1699 | 0.0866 | 0.1149 | 0.2231 | 0.1680 | 0.1683 | | 0.2546 |
| NuminaMath | OPEN-ENDED | 0.1484 | 0.2943 | 0.1004 | 0.1949 | 0.1516 | 0.2020 | 0.4460 | |
| CrossThink-MATH | OPEN-ENDED | 0.1699 | 0.3370 | 0.1149 | 0.2231 | 0.1736 | 0.1579 | 0.5105 | |
| MATH | OPEN-ENDED | 0.0145 | 0.0287 | 0.0098 | 0.0190 | 0.0148 | 0.1691 | 0.0435 | |

Table 11: Proportion of each dataset in different blends.

| Task Type | Model | Min | Max | Mean |
|---|---|---|---|---|
| GPR | $\mathcal{B}_{gpr\uparrow}$ | 83.20 | 2697.80 | 385.41 |
| | $\mathcal{B}_{only\_mr}$ | 159.60 | 9594.00 | 638.57 |
| | ORZ | 223.00 | 8221.80 | 1114.60 |
| Math | $\mathcal{B}_{gpr\uparrow}$ | 170.25 | 10130.00 | 622.00 |
| | $\mathcal{B}_{only\_mr}$ | 201.75 | 11330.25 | 730.68 |
| | ORZ | 292.00 | 12917.00 | 1257.00 |

Table 12: Token length statistics (Min, Max, Mean) for correct responses across task types.

multiple choice and proof-oriented problems which reduces the answer space diversity. MCQ type of questions are important for MMLU and other non-reasoning centric tasks. Recently, (Ma et al., 2025; Su et al., 2025) attempted to address this with a model-based verifier to handle diversity in the answer space. However, as discussed in previous works, LLM-as-a-Judge may suffer from pitfalls of reward hacking (DeepSeek-AI, 2025; Weng, 2024; Wen et al., 2024) and further diverge the model more from the correct reasoning processes. Additionally, they do not show any analysis to estimate the contribution of each domain in the final task performance. Despite training on diverse domains, CrossThink offers simple, scalable and robust reward estimation without any external reward model. For a rule-based reward model, the format of input data and the final answer is crucial and largely underexplored. Furthermore, their additional sources of data synthesis approach has no details making it infeasible to scale for domains other than math. The kind of data and the ratio of each type of data important for the overall improvement of LLMs across multiple benchmarks have yet to be explored.

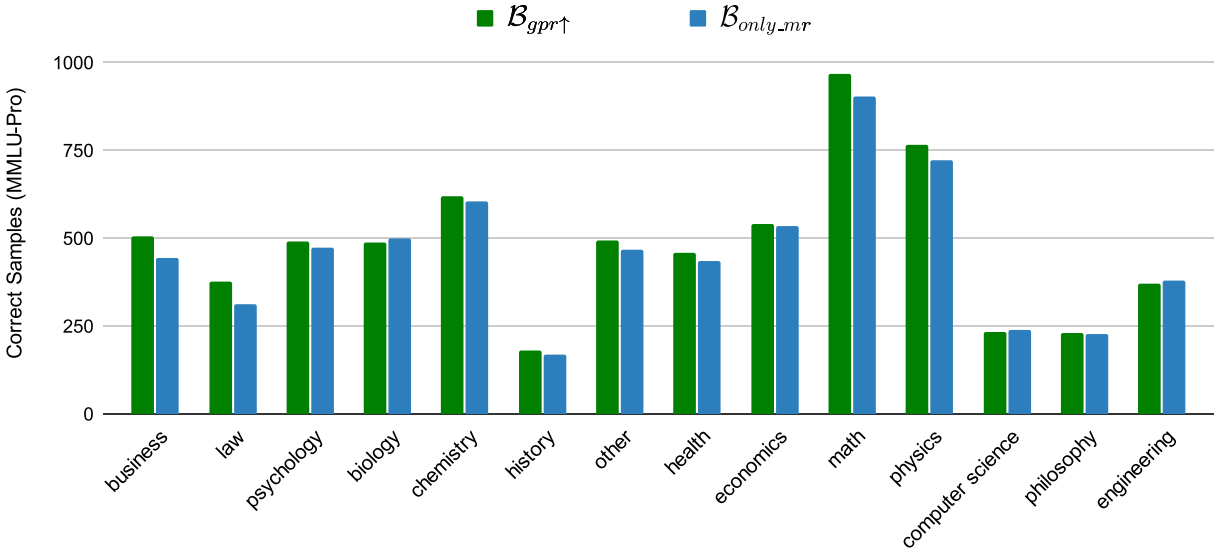


Figure 8: **Sub-category Accuracy Comparison across MMLU-PRO Domains.** The $\mathcal{B}_{gpr\uparrow}$ blend consistently outperforms $\mathcal{B}_{only\_mr}$ in a wide range of non-math reasoning categories such as business, law, psychology, and economics. Surprisingly, it also slightly surpasses the math-specialized blend in the MMLU-PRO math category, highlighting the generalizability and versatility of multi-domain training.

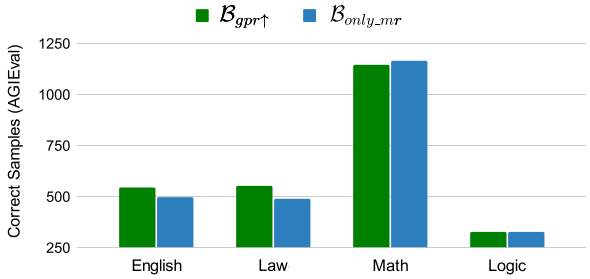


Figure 9: **Sub-category Accuracy Comparison across AGIEVAL.** While $\mathcal{B}_{only\_mr}$ performs marginally better in the math, $\mathcal{B}_{gpr\uparrow}$ achieves stronger results in non-math domains.

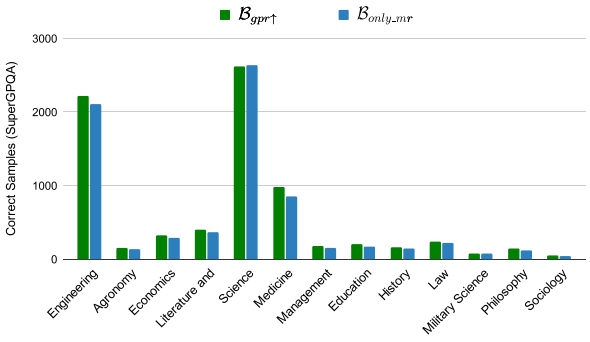


Figure 10: **Sub-category Accuracy Comparison across SUPERGPQA.** The $\mathcal{B}_{gpr\uparrow}$ blend consistently outperforms $\mathcal{B}_{only\_mr}$ in a wide range of non-math reasoning categories except the science category which consists of fields like mathematics, physics, astronomy, chemistry etc.—highlighting the generalizability and versatility of multi-domain training.