MEML-GRPO: Heterogeneous Multi-Expert Mutual Learning for RLVR Advancement

Weitao Jia[1]*, Jinghui Lu[1]*†, Haiyang Yu[1 2]*, Siqi Wang[1], Guozhi Tang[1], An-Lan Wang[1], Weijie Yin[1], Dingkang Yang[1], Yuxiang Nie[1], Bin Shan[1], Hao Feng[1], Irene Li[3], Kun Yang[2], Han Wang[1], Jingqun Tang[1], Teng Fu[2], Changhong Jin[4], Chao Feng[1]†, Xiaohui Lv[1]†, Can Huang[1]†

[1] ByteDance Inc.
[2] Fudan University
[3] University of Tokyo
[4] University College Dublin

Abstract

Recent advances demonstrate that reinforcement learning with verifiable rewards (RLVR) significantly enhances the reasoning capabilities of large language models (LLMs). However, standard RLVR faces challenges with reward sparsity, where zero rewards from consistently incorrect candidate answers provide no learning signal, particularly in challenging tasks. To address this,we propose **M**ulti-**E**xpert **M**utual **L**earning **GRPO** (MEML-GRPO), an innovative framework that utilizes diverse expert prompts as system prompts to generate a broader range of responses, substantially increasing the likelihood of identifying correct solutions. Additionally, we introduce an inter-expert mutual learning mechanism that facilitates knowledge sharing and transfer among experts, further boosting the model's performance through RLVR. Extensive experiments across multiple reasoning benchmarks show that MEML-GRPO delivers significant improvements, achieving an average performance gain of 4.89% with Qwen and 11.33% with Llama, effectively overcoming the core limitations of traditional RLVR methods.

1 Introduction

Recent advances in large language model (LLM) and reasoning (Kojima et al. 2022; Wei et al. 2022; Wang et al. 2022; Lyu et al. 2023; Feng et al. 2023; Shah et al. 2024; Zhang et al. 2024; Yu et al. 2023; Wang et al. 2025a; Zhang et al. 2025; Zhou, Luo, and Jiang 2025; Lu et al. 2025a,b; Yu et al. 2025; Lu et al. 2024, 2023a,b, 2022) have showcased the efficacy of reinforcement learning with verifiable rewards (RLVR) in improving reasoning capabilities. Models such as OpenAI-o1 (Jaech et al. 2024), DeepSeek-r1 (Shao et al. 2024; Guo et al. 2025), Doubao-1.5-thinking (Team 2025), and Qwen QwQ (Team 2024) demonstrate the effectiveness of this approach in enhancing logical and problem-solving performance. By optimizing models based on binary correctness signals—such as matching ground truth solutions (Xia et al. 2025) in mathematics or passing unit tests in code (Lai et al. 2025)—RLVR provides a scalable and automated approach to improving reasoning without relying on expensive human annotations (Ouyang et al. 2022; Tong et al. 2024) or Process Reward Models (Setlur et al. 2024).

Despite its successes, some studies (Yue et al. 2025; Shojaee et al. 2025; Zhao et al. 2025) argue that RLVR does not endow LLMs with genuinely new reasoning abilities beyond those already present in their base models. Instead, RLVR primarily improves performance by steering models toward reasoning paths they already know, which are more likely to yield rewards. In other words, most RLVR methods optimize within the model's existing knowledge rather than enabling the acquisition of new information. This limitation is most evident in what is termed *reward sparsity*: when a model's initial policy fails to produce correct responses in complex reasoning tasks, the lack of positive learning signals prevents RLVR from driving meaningful progress. This limitation becomes particularly pronounced in scenarios requiring exploration beyond the model's current knowledge. The on-policy nature of current RLVR methods, which rely solely on a model's own generated trajectories for learning, inherently limits their ability to explore beyond existing capabilities. When a model's initial reasoning capacity is inadequate, standard RLVR tends to plateau early. This constraint, closely tied to the issue of reward sparsity described earlier, highlights a critical challenge: *How can RLVR be designed to overcome reward sparsity and uncover correct reasoning paths, even when the initial policy falls short?*

To overcome this challenge, we introduce Heterogeneous **M**ulti-**E**xpert **M**utual **L**earning **GRPO** (MEML-GRPO), an innovative framework that leverages the complementary strengths of multiple heterogeneous pre-trained models to overcome performance bottlenecks. Our approach is inspired by the observation that diverse, heterogeneous reasoning models often generate varied and complementary solutions to the same problem. As shown in Table 1, error distributions in GSM8K (Cobbe et al. 2021) across these models (*i.e.,* DeepSeek-r1, GPT4o[1] and Doubao-1.5-thinking) exhibit low overlap, with only 3.06% of errors shared among

* Equal contribution.

† Corresponding author.

[1]Model card:gpt-4o-2024-05-13

all three models. Notably, in the GSM8K dataset, 90.11% of incorrect predictions from GPT4o can be corrected by other heterogeneous models, highlighting the potential for mutual learning to enhance overall performance.

By fine-tuning the policy model with responses from diverse models, MEML-GRPO boosts the likelihood of producing at least one valid reasoning path, delivering the essential learning signal for RLVR. Moreover, instead of merely aggregating the learned outputs of the policy model from these heterogeneous models, MEML-GRPO enables these models to learn from each other's strengths and promote mutual improvement. MEML-GRPO introduces three key innovations to enhance reasoning in RLVR:

1. **Multi-Expert Fine-Tuning (MEF)** (Section 3.1): Instead of relying on a single reasoning dataset, MEML-GRPO utilizes multiple system prompts, referred to as *experts* in this work, each emulating distinct reasoning paths from diverse external models. These experts are fine-tuned on responses generated by heterogeneous models, with their unique reasoning styles captured through tailored system prompts, which bring a diverse set of reasonings that enrich the learning process.
2. **Reinforced Inter-Expert Learning (RIEL)** (Section 3.2): Beyond independent exploration, MEML-GRPO enables experts to learn from one another's successful reasoning paths through a shared mechanism. Weaker experts improve by learning high-reward trajectories from stronger ones. As a result, all experts can perform competitively majority voting during inference.
3. **Hard Example Accumulation via SFT Buffer** (Section 3.2): When all experts fail to solve a problem, MEML-GRPO defaults to supervised fine-tuning (SFT) using ground truth, ensuring non-zero gradients for learning, maintaining progress even on challenging problems.

MEML-GRPO strikes an optimal balance between exploration and exploitation. Unlike traditional RLVR methods that may stall on difficult problems, our framework dynamically integrates complementary strengths from multiple reasoning strategies, ensuring steady learning progress. Our contributions are:

- We propose MEML-GRPO, which leverages system prompts to capture diverse reasoning styles and introduces a reinforced mutual learning algorithm for cross-expert knowledge transfer.
- Extensive experiments on GSM8K, MathQA, and StrategyQA demonstrate consistent performance improvements of 3.08%, 8.65%, and 2.96% with Qwen and 5.87%, 16.78%, and 11.35% with Llama, respectively, over state-of-the-art (SOTA) RLVR methods, demonstrating MEML-GRPO's ability to address the reward sparsity problem in RLVR.

2 Related Work

Recent advances in reinforcement learning (RL) have significantly enhanced the reasoning capabilities of large language models (Kojima et al. 2022; Wang et al. 2022; Niu et al. 2025b,a), as demonstrated by models such as DeepSeek-r1 (Guo et al. 2025), OpenAI-o1 (Jaech et al. 2024), Doubao-1.5-thinking (Team 2025). Notably, RL with verifiable rewards (RLVR) (Wang et al. 2025b) has been systematically studied, revealing its effectiveness in enabling complex reasoning. Given the significant performance improvements achieved by RLVR, there has been growing research interest in further advancing its training procedure. For example, Muennighoff et al. (2025); Li et al. (2025) demonstrate promising improvements in RL through test-time computation scaling, however, their experimental results reveal that performance remains constrained by the model's inherent knowledge. Ye et al. (2025); Fatemi et al. (2025) demonstrate that introducing structured CoT reasoning paths can elicit advanced reasoning abilities. Su et al. (2025); Wen et al. (2025) introduce novel training paradigms specifically designed to improve reasoning. However, recent studies (Yue et al. 2025) reveal that RLVR's on-policy learning faces difficulties in exploration spaces, as it predominantly focuses on biasing the model toward behaviors that are more likely to yield rewards instead of learning new knowledge or reasoning paths. As a result, these methods tend to exploit familiar reasoning patterns rather than explore new knowledge-based reasoning paths. To address this challenge, we propose MEML-GRPO that leverages divers reasoning paths generated from heterogeneous pre-trained models to transcend these cognitive constraints while preserving self-driven exploration capabilities.

3 Methodology

The proposed framework consists of two main stages: Multi-Expert Fine-tuning (MEF) and Reinforced Inter-Expert Learning (RIEL). The former aims to equip a base language model with the ability to emulate multiple expert behaviors, while the latter further enhances the model's performance through mutual learning among experts using reinforcement learning techniques.

3.1 Multi-Expert Fine-tuning (MEF)

The MEF stage is designed to endow the base model with multi-expert capabilities by fine-tuning it on a dataset that contains answers generated by various pre-trained heterogeneous LLMs under different expert prompts. This allows the model to learn how to produce distinct responses conditioned on specific expert instructions.

Dataset Construction via Expert Prompting. Let $\mathcal{E} = \{E_1, E_2, ..., E_N\}$ denote the set of N pre-trained heterogeneous expert models. For each question Q in our training set $\mathcal{Q}$, we generate an answer from each expert E_i:

$$A^{(i)} = E_i(Q) \tag{1}$$

To construct the expert-specific samples $\mathcal{D}_{\text{ME}}$, we supplement corresponding prompt after the question Q. This results in a multi-expert dataset:

$$\mathcal{D}_{\text{ME}} = \left\{ (\texttt{Concat}(Q_j, P_i), A_j^{(i)}) \mid Q_j \in \mathcal{Q},\ i = 1, ..., N \right\} \tag{2}$$

where $A_j^{(i)}$ is the answer given by expert i to question j. P_i represents the control instruction of the i-th expert such as:

Metric	DeepSeek-r1	GPT4o	Doubao-1.5-thinking	Error Overlap
GSM8K				
Total errors	748 (10.0%)	2316 (30.9%)	388 (5.1%)	229 (3.0%)
Errors corrected by other models	519 (69.3%)	2087 (90.1%)	159 (40.9%)	–
StrategyQA				
Total errors	400 (20.0%)	326 (16.3%)	369 (18.4%)	192 (9.6%)
Errors corrected by other models	208 (52.0%)	134 (41.1%)	177 (47.9%)	–
MathQA				
Total errors	3893 (13.0%)	5004 (16.7%)	3164 (10.6%)	2281 (7.6%)
Errors corrected by other models	1612 (41.4%)	2723 (54.4%)	863 (27.9%)	–

Table 1: Error distribution analysis across heterogeneous models on different datasets. Percentages show error rates (total errors/examples) and correction rates (errors fixed by other models/total errors). Error overlap indicates shared errors.

*"You are **Expert** i, please provide an answer to the above question."*. The function $\texttt{Concat}(\cdot)$ is used to concatenate Q and P. A more detailed prompt would be: *"You are Expert DeepSeek. [Q]"* where [Q] is the specific question.

Fine-tuning Procedure. We adopt a strong base LLM and fine-tune it using the constructed dataset $\mathcal{D}_{\text{ME}}$. The objective is to maximize the conditional log-likelihood of the expert answers given their respective prompts:

$$\mathcal{L}_{\text{MEF}} = -\sum_{j=1}^{M}\sum_{i=1}^{N} \log p_\theta \left(A_j^{(i)} \mid Q_j, \mathrm{P}_i\right) \tag{3}$$

where θ denotes the parameters of the adopted LLM, and M is the number of questions in $\mathcal{Q}$. After this stage, the model becomes capable of generating diverse expert-like responses depending on the input prompt. That is, for any question Q, the model can be instructed to respond like expert i by simply add the corresponding prompt P_i to Q.

3.2 Reinforced Inter-Expert Learning (RIEL)

While the MEF stage equips the model with expert knowledge, the RIEL stage further improves its reasoning and exploration capabilities through reinforcement learning and inter-expert knowledge transfer.

Response Sampling and Intra-Expert Advantage Estimation. For a given question Q, we sample G responses from each expert policy induced by the prompt prompt_i:

$$\mathcal{O}(Q) = \left\{\left\{O_1^1, ..., O_G^1\right\}, ..., \left\{O_1^N, ..., O_G^N\right\}\right\} \tag{4}$$

Each response $O_g^i \sim \pi_\theta(\cdot \mid Q, P_i)$ is generated from the current policy model parameterized by θ. We then compute a reward function $r(O_g^i)$ to evaluate the quality of each response. For example, the reward can be derived from task-specific metrics, such as accuracy, or rule-based criteria, such as matching a regular expression format.

In each expert group i, we compute the advantage of each response by comparing it to the average reward for that expert, following the approach in GRPO (Shao et al. 2024):

$$A_g^i = r(O_g^i) - \frac{1}{G}\sum_{g'=1}^{G} r(O_{g'}^i) \tag{5}$$

Using these advantages, we extend the GRPO loss for expert i as follows:

$$\mathcal{L}_{\text{GRPO}}^{(i)} = -\mathbb{E}_{O_g^i \sim \pi_\theta}\left[\log \pi_\theta(O_g^i \mid Q, P_i) \cdot \max\left(A_g^i, 0\right)\right] \tag{6}$$

The total GRPO loss across all experts is:

$$\mathcal{L}_{\text{GRPO}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{GRPO}}^{(i)} \tag{7}$$

This encourages the model to reinforce responses that outperform the average within the same expert group.

Inter-Expert Mutual Learning via KL Divergence Regularization. To promote knowledge exchange between experts, we introduce a mechanism called inter-expert mutual learning. For each question Q, we identify the best-performing expert E^+ and the worst-performing expert E^- based on their average reward:

$$\begin{aligned} E^+ &= \arg\max_{i \in \{1,\ldots,N\}} \frac{1}{G}\sum_{g=1}^{G} r(O_g^i), \\ E^- &= \arg\min_{i \in \{1,\ldots,N\}} \frac{1}{G}\sum_{g=1}^{G} r(O_g^i) \end{aligned} \tag{8}$$

We define $prompt_{E^-}$ as the system prompt of expert E^-, $prompt_{E^+}$ as the system prompt of expert E^+. O^+ is the correct responses generated by E^+. Thus, $\log p_\theta(O^+ \mid Q, \text{prompt}_{E^-})$ denotes the log-probability of the high-quality response from E^-, and $\log p_\theta(O^+ \mid Q, \text{prompt}_{E^+})$ is the log-probability of the high-quality response from E^+.

We implement the KL divergence as a loss penalty to enable the less effective expert to learn from the more effective expert, as follows:

$$\begin{aligned} \mathcal{L}_{\text{KL}} \approx &\log p_\theta(O^+ \mid Q, \text{prompt}_{E^-}) \\ &- \log p_\theta(O^+ \mid Q, \text{prompt}_{E^+}) \end{aligned} \tag{9}$$

The KL divergence-based regularization term penalizes differences between the output distributions of the poorly performing expert (E^-) and the well-performing expert (E^+), encouraging the former to adopt the latter's strengths. A key benefit of this inter-expert mutual learning approach

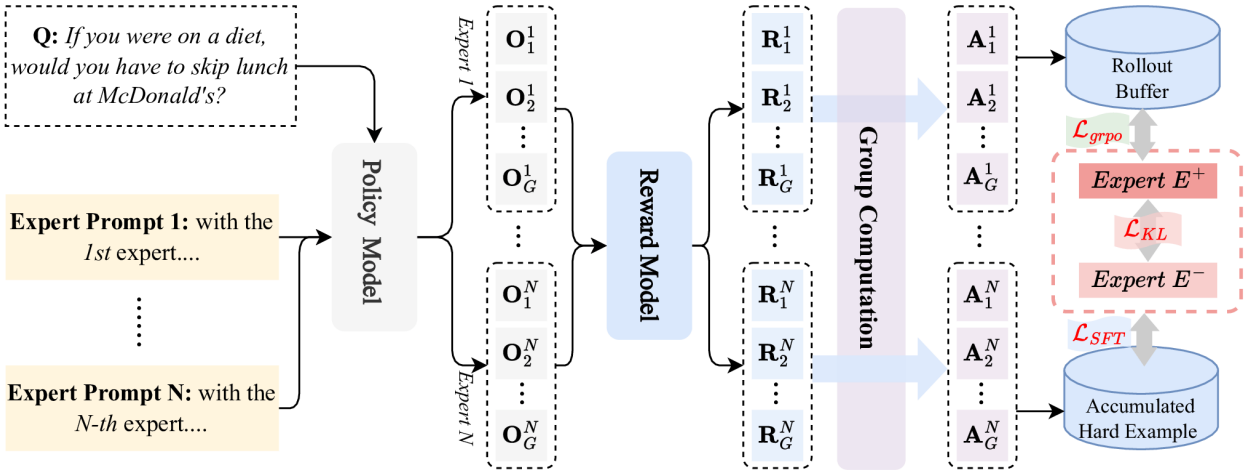


Figure 1: This figure illustrates the pipeline of MEML-GRPO. The GRPO loss, which is computed across all experts, is omitted from the figure for brevity.

is its efficiency during inference. Unlike conventional methods like majority voting, which require multiple inferences per question (proportional to the number of experts), mutual learning enables each expert to absorb the advantages of others during training. Thus at inference time, we can select the best-performing expert, significantly reducing computational costs while maintaining high-quality responses.

Hard Example Accumulation via SFT Buffer. Despite the benefits of inter-expert learning, there is a risk of error propagation when the top-performing expert also generates incorrect responses. To mitigate this issue, we design a Hard Example Accumulation Mechanism. We maintain a buffer $\mathcal{B}$ of capacity B to store difficult samples. For each question Q, if expert i produces more than K incorrect answers out of G responses, we add the pair $(Q, P_i) \rightarrow O_{\text{gt}}$ into the buffer with probability $\frac{K}{G}$, where O_{gt} is the correct answer. Once the buffer reaches full capacity, we perform supervised fine-tuning periodically during:

$$\mathcal{L}_{\text{SFT}} = -\sum_{(Q,P_i,O_{\text{gt}})\in\mathcal{B}} \log p_\theta(O_{\text{gt}} \mid Q, P_i) \qquad (10)$$

In this work, the buffer capacity ($\mathcal{B}$) is set to 64. A question is flagged as a hard example and added to the buffer with a probability of $K/G = 75\%$ if experts produce more than $K = 6$ incorrect answers out of $G = 8$ samples. An SFT update is triggered only when the buffer is full. Therefore, on a challenging dataset, the buffer fills more rapidly, resulting in more frequent SFT updates. This ensures the model continues to learn from difficult cases.

3.3 Overall Training Objective

The overall training objective combines all components introduced above:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \lambda_1 \mathcal{L}_{\text{KL}} + \lambda_2 \mathcal{L}_{\text{SFT}} \qquad (11)$$

Model	Dataset			
	GSM8K	StrategyQA	MathQA	Avg.
Qwen2.5-1.5B-Math				
Expert0-SFT	**67.0**	**69.6**	68.0	**68.2**
Expert1-SFT	48.5	66.0	51.0	55.1
Expert2-SFT	66.0	67.2	65.0	66.0
MoE-SFT (Expert0)	**67.0**	67.0	**70.0**	68.0
MoE-SFT (Expert1)	48.0	68.0	53.0	56.3
MoE-SFT (Expert2)	**67.0**	65.0	65.0	65.6
Llama3.2-1B-Instruct				
Expert0-SFT	**45.0**	55.0	56.0	52.0
Expert1-SFT	35.0	55.0	**56.2**	48.7
Expert2-SFT	42.0	57.0	48.0	49.0
MoE-SFT (Expert0)	42.6	59.3	54.9	**52.2**
MoE-SFT (Expert1)	36.2	**65.8**	46.0	49.3
MoE-SFT (Expert2)	36.7	63.0	52.0	50.5

Table 2: Accuracy of MoE-SFT compared to individual Expert SFT methods. Expert0 refers to ground truth, Expert1 to DeepSeek-r1, and Expert2 to Doubao-1.5-thinking.

where λ_1 and λ_2 are hyperparameters balancing the contributions of each loss component.

4 Experimental Results

4.1 Experimental Settings

Datasets. We evaluate MEML-GRPO on text datasets, selecting three widely used reasoning datasets: the mathematical reasoning datasets **GSM8K** (Cobbe et al. 2021) and **MathQA** (Amini et al. 2019), as well as the commonsense reasoning dataset **StrategyQA** (Geva et al. 2021). Besides ground truth reasoning paths (*i.e.,* **Expert0**), other off-policy reasoning trajectories are generated by DeepSeek-r1 (*i.e.,* **Expert1**), and Doubao-1.5-thinking (*i.e.,* **Expert2**).

Item	Description/Reasoning	Final Answer	Correctness
Problem	Jenny has a pizza with 12 slices. She gives $\frac{1}{3}$ to Bill and $\frac{1}{4}$ to Mark. After eating 2 slices herself, how many slices are left?	–	–
Expert0	Calculates total given: $\frac{1}{3}+\frac{1}{4}=\frac{4}{12}+\frac{3}{12}=\frac{7}{12}$ (7 slices). Jenny eats 2 slices, leaving $12-2=10$. Subtracts given slices: $10-7=3$.	3	Yes
Expert1	Gives Bill $12\times\frac{1}{3}=4$ slices, Mark $12\times\frac{1}{4}=3$ slices. Jenny eats 2 slices. Calculates: $12-4-3-2=5$.	5	No
Expert2	Calculates total given: $\frac{1}{3}=\frac{4}{12}$, $\frac{1}{4}=\frac{3}{12}$, so $\frac{4}{12}+\frac{3}{12}=\frac{7}{12}$ (7 slices). Subtracts only Jenny's 2 slices: $12-2=10$.	10	No
GT	-	3	–

Table 3: Illustration of different reasoning paths generated by different system prompts. Expert0 refers to ground truth, Expert1 to DeepSeek-r1, and Expert2 to Doubao-1.5-thinking.

Evaluation Metric. For all datasets, as in Touvron et al. (2023); Wang et al. (2024); Qwen (2023), We use exact-match accuracy to determine correctness.

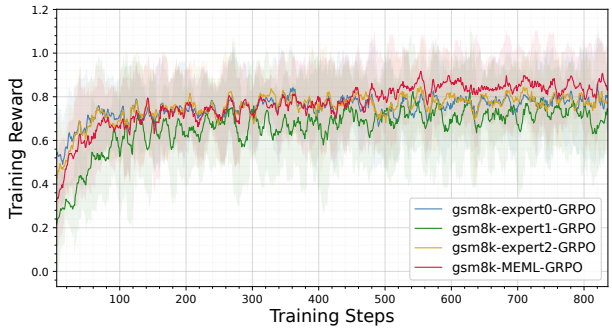
Baselines. For SFT methods, we consider the following methods: (1) **Expert$X$-SFT**: models fine-tuned exclusively on specific reasoning trajectories, namely **Expert0** (ground truth), **Expert1** (DeepSeek-r1), and **Expert2** (Doubao-1.5-thinking); (2) **MoE-SFT**: a single model trained on all reasoning trajectories using system prompt to distinguish reasoning style as described in Section 3.1. For RLVR methods, we evaluate the following approaches: (1) **Expert$X$-GRPO**: models fine-tuned on specific reasoning trajectories and then trained using GRPO (Guo et al. 2025); (2) **Expert$X$-Dr.GRPO**: models fine-tuned on specific reasoning trajectories and then trained using Dr.GRPO (Liu et al. 2025), an advanced variant of GRPO serving as a stronger baseline; (3) **MoE-SFT-GRPO**: models fine-tuned on all reasoning trajectories and then trained using GRPO; (4) **MoE-SFT-Dr.GRPO**: models fine-tuned on all reasoning trajectories and then trained using Dr.GRPO; (5) **MEML-GRPO**: models fine-tuned on all reasoning trajectories and then trained using the proposed MEML-GRPO method.

Training and Inference Setup. We conduct experiments on both Qwen2.5-1.5B-Math (Yang et al. 2024), Llama3.2-1B-Instruct (Grattafiori et al. 2024) to verify the generalization of MEML-GRPO. To ensure fairness, we maintain 8 rollouts per prompt for all RL-trained models. The learning rate is set to 1×10^{-6}. All training experiments are conducted on 8 A800 GPUs. SFT training setup, The learning rate is set to 1×10^{-5}. We train all RL models for 1 epoch and all SFT models for 1 epochs. For inference, to eliminate the impact of randomness, no sampling methods are employed during testing for any of the models. Greedy search is used for generation across all models.

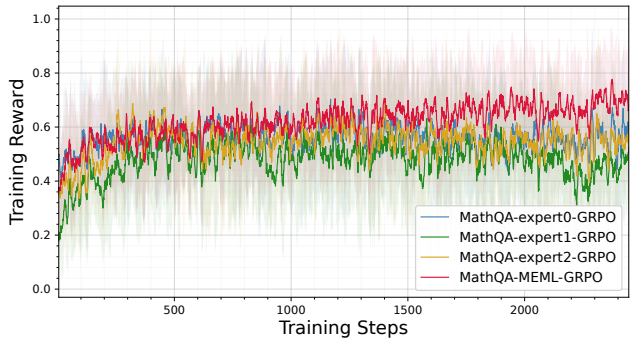
4.2 Main Results

Multi-expert Learning for SFT. As presented in Table 2, for instance, in the GSM8K dataset with Llama3.2 the three experts in MoE-SFT achieve scores of 42.6%, 36.2%, and 36.7%, respectively. These results are slightly lower than those of individually trained Expert0-SFT (45.0%) and Expert2-SFT (42.0%). A similar trend is observed in MathQA and StrategyQA, as well as with Qwen2.5. The notable exception is StrategyQA with Llama3.2 where one expert achieves a score of 65.8%, significantly surpassing the performance of individually trained SFT experts. These findings suggest that MoE-SFT can effectively leverage the knowledge of multiple experts to some degree, with each system prompt enabling the same model to function as distinct experts, reflecting the reasoning behaviors of corresponding heterogeneous models. However, the absence of an effective mechanism to fully utilize the knowledge of multiple experts limits MoE-SFT's ability to consistently outperform individually trained expert SFT models across all scenarios, highlighting the effectiveness of MEML-GRPO.

Comparison with SOTA RLVR methods. Table 4 compares the performance of MEML-GRPO with other RL baselines. MEML-GRPO consistently achieves top performance across all datasets for both Qwen2.5 and Llama3.2 models. Note that to ensure a fair comparative evaluation, we partition the dataset into two distinct subsets: 20% for warm-up SFT and the remaining 80% for RL optimization, differing from the setup used in Table 2. We first compare MEML-GRPO with the single-expert RL method, Expert0-SFT-GRPO, which serves as the standard GRPO baseline since Expert0 is fine-tuned with ground-truth reasoning data. Using Qwen2.5, MEML-GRPO(Expert0) outperforms this baseline by significant margins of 4.2%, 14.5%, and 5.1% across the datasets. Similarly, when compared to Expert0-SFT-Dr.GRPO, a standard Dr.GRPO method, MEML-GRPO demonstrates substantial improvements. Consistent trends are observed with Llama3.2, highlighting the robustness of MEML-GRPO across different models. Note that the single-expert RL method, when trained with data generated by other experts, also performs less effectively than the MEML-GRPO with the corresponding expert.



(a) GSM8K



(b) MathQA

(c) StrategyQA

Figure 2: Training reward dynamics of MEML-GRPO (Llama3.2) compared with other on-policy RL.

Model	Dataset			
	GSM8K	StrategyQA	MathQA	Avg.
Qwen2.5-1.5B-Math				
Single Expert				
Expert0-SFT-GRPO	75.9	58.6	71.3	67.1
Expert1-SFT-GRPO	73.1	62.5	71.8	69.1
Expert2-SFT-GRPO	76.2	60.4	69.3	68.6
Expert0-SFT-Dr.GRPO	78.3	55.2	72.4	68.6
Expert1-SFT-Dr.GRPO	76.1	64.5	73.3	71.3
Expert2-SFT-Dr.GRPO	77.2	62.4	70.6	70.0
Multiple Expert				
MoE-SFT-GRPO (Expert0)	77.1	68.2	70.4	71.9
MoE-SFT-GRPO (Expert1)	75.6	70.3	68.5	71.4
MoE-SFT-GRPO (Expert2)	74.3	67.9	70.4	70.8
MoE-SFT-Dr.GRPO (Expert0)	77.3	68.6	72.1	72.6
MoE-SFT-Dr.GRPO (Expert1)	77.7	69.3	74.2	73.7
MoE-SFT-Dr.GRPO (Expert2)	77.6	68.7	70.1	72.1
Ours				
MEML-GRPO (Expert0)	**80.1**	73.1	**76.4**	76.5
MEML-GRPO (Expert1)	79.6	**75.3**	**76.4**	**77.1**
MEML-GRPO (Expert2)	78.1	74.1	76.3	76.2
Llama3.2-1B-Instruct				
Single Expert				
Expert0-SFT-GRPO	56.3	54.1	52.5	54.3
Expert1-SFT-GRPO	58.2	54.0	45.2	52.4
Expert2-SFT-GRPO	54.5	53.1	47.9	51.8
Expert0-SFT-Dr.GRPO	57.3	55.6	54.3	55.7
Expert1-SFT-Dr.GRPO	58.4	54.8	46.7	53.3
Expert2-SFT-Dr.GRPO	54.5	53.7	48.5	52.2
Multiple Expert				
MoE-SFT-GRPO (Expert0)	57.7	53.7	55.1	55.5
MoE-SFT-GRPO (Expert1)	57.1	53.8	55.2	55.3
MoE-SFT-GRPO (Expert2)	56.7	53.4	54.0	54.7
MoE-SFT-Dr.GRPO (Expert0)	57.9	53.1	54.6	55.2
MoE-SFT-Dr.GRPO (Expert1)	57.2	54.2	54.2	55.2
MoE-SFT-Dr.GRPO (Expert2)	57.4	53.1	53.9	54.8
Ours				
MEML-GRPO (Expert0)	60.3	62.0	**60.8**	**61.1**
MEML-GRPO (Expert1)	61.0	**63.3**	58.4	60.9
MEML-GRPO (Expert2)	**61.3**	61.7	58.8	60.6

Table 4: Accuracy of MEML-GRPO compared to other RL baselines. Expert0 refers to ground truth, Expert1 to DeepSeek-r1, and Expert2 to Doubao-1.5-thinking.

Comparison with Other Multi-expert RLVR Methods. By integrating responses from all experts in the training set, standard GRPO and Dr.GRPO can be extended to Multi-expert RLVR methods, namely MoE-SFT-GRPO and MoE-SFT-Dr.GRPO, establishing stronger baselines. Table 4 shows that the MEML-GRPO method substantially outperforms state-of-the-art approaches across all reasoning tasks, delivering consistent improvements with both Qwen2.5 and Llama3.2 models, as evidenced by its superior overall performance. For example, on the Qwen2.5-1.5B-Math model, MEML-GRPO achieves an average accuracy of 79.3% on the GSM8K dataset, surpassing its GRPO and Dr.GRPO counterparts. Notably, on StrategyQA, the three trained experts under MEML-GRPO achieve accuracies of 73.1%, 75.3%, and 74.1%, substantially outperforming GRPO (68.2%, 70.3%, 67.9%) and Dr.GRPO (68.6%, 69.3%, 68.7%). Similarly, with Llama3.2, MEML-GRPO also demonstrates superior performance across all three datasets.

The possible reasons for the performance differences mainly lie in the two core designs of MEML-GRPO: (1) The multi-expert prompt mechanism generates more diverse responses through varied system prompts, significantly increasing the probability of covering correct solutions. This

			Qwen2.5-1.5B-Math				Llama3.2-1.5B-Instruct			
MoE	HSFT	IML	GSM8K	StrategyQA	MathQA	Avg	GSM8K	StrategyQA	MathQA	Avg
×	×	×	76.2	62.5	71.8	70.1	56.3	54.1	52.5	54.3
✓	×	×	77.1	70.3	70.4	72.6	57.7	53.8	55.2	55.5
✓	✓	×	79.9	74.3	75.1	76.4	59.4	61.2	58.8	59.8
✓	×	✓	78.1	72.3	73.4	74.6	58.4	56.8	57.3	57.5
✓	✓	✓	**80.1**	**75.3**	**76.4**	**77.3**	**61.3**	**63.3**	**60.8**	**61.8**

Table 5: Ablations on each component of MEML-GRPO. For brevity, we report the results of best-performing experts. MoE: multiple expert SFT. HSFT: hard example SFT. IML: Inter-expert mutual learning.

Model	Dataset			
	GSM8K	StrategyQA	MathQA	Avg.
Qwen2.5-1.5B-Math				
Multiple Expert				
MoE-SFT-GRPO (Expert0)	77.1	68.2	70.4	71.9
MoE-SFT-GRPO (Expert1)	75.6	70.3	68.5	71.4
MoE-SFT-GRPO (Expert2)	74.3	67.9	70.4	70.8
MoE-SFT-GRPO (MV)	**78.3**	**71.1**	**71.6**	**73.6**
Delta	1.2	0.8	1.2	1.1
Ours				
MEML-GRPO (Expert0)	**80.1**	73.1	**76.4**	76.6
MEML-GRPO (Expert1)	79.6	**75.3**	**76.4**	**77.1**
MEML-GRPO (Expert2)	78.1	74.1	76.3	76.2
MEML-GRPO (MV)	79.8	75.1	75.9	76.9
Delta	-0.3	-0.2	-0.5	-0.3

Table 6: Comparison of majority voting performance versus a single expert, with delta calculated as the difference between majority vote performance and the best-performing expert. MV: majority voting.

particularly alleviates the reward sparsity problem in traditional RLVR where "no learning signal is provided when all candidate answers are wrong" in complex reasoning tasks. (2) The inter-expert mutual learning mechanism facilitates knowledge sharing and transfer among experts, further breaking through the capability ceiling of the model through reinforcement learning. In contrast, MoE-SFT-GRPO and MoE-SFT-Dr.GRPO only rely on a single reinforcement learning strategy. They lack active enhancement of response diversity and collaborative learning among experts, thus failing to fully exploit the model's potential in reasoning tasks and resulting in limited performance.

Table 3 presents a qualitative comparison of reasoning paths provided by different experts (*i.e.,* system prompts) for a pizza slice distribution problem that is sampled from GSM8K. Figure 2 shows the training reward dynamics of MEML-GRPO (Llama3.2) compared with other on-policy RL method, the results with Qwen2.5 are in Appendix.

4.3 Discussion

On the Benifit of Mutual Learning. Table 6 presents the distinct differences in delta values, demonstrating that our MEML-GRPO models outperform the majority vote (MV) approach across all datasets. Compared to majority voting, MEML-GRPO achieves higher accuracy, as evidenced by the delta values. This improvement highlights the effectiveness of mutual learning in transferring knowledge between different experts, enabling the reasoning results to rival or even match the performance of multi-expert majority voting. In contrast, conventional schemes typically achieve their best results with majority voting. The results with Llama3.2 can be found in Appendix.

Ablation Study. Table 5 shows the impact of individual components in the MEML-GRPO framework. The baseline achieved average scores of 70.1% for Qwen2.5 and 54.3% for Llama3.2. Enabling Mixture of Experts (MoE) alone improved the averages to 72.6% and 55.5%, respectively. Adding Hypothesis Selection Fine-Tuning (HSFT) further increased the scores to 76.4% and 59.8%. Incorporating Interactive Multi-step Learning (IML) alone provided an average gain of 2% for both models. The full configuration (MoE+HSFT+IML) yielded the highest averages: 77.3% for Qwen2.5 and 61.8% for Llama3.2, showing that each component contributes and their combination is most effective.

Training and Inference Cost MEML-GRPO achieves remarkable training efficiency, with a total cost significantly lower than expected. By utilizing the vLLM inference engine for parallel batched rollouts and paged attention, the total training time increases by only 20-30% over the single-expert baseline—far less than the theoretical N-fold increase (where N=3, the number of experts). Furthermore, MEML-GRPO's peak memory footprint is comparable to conventional GRPO (around 60GB), as vLLM pre-allocates memory based on inference settings. This efficiency extends to inference: after training, only a single model is deployed. By dynamically selecting the best expert prompt, MEML-GRPO delivers superior performance with the latency of a single model, avoiding the computational overhead of ensemble methods that require running N models per input.

5 Conclusion

In this work, we introduce MEML-GRPO, a multi-expert mutual learning framework that mitigates reward sparsity in RL for LLM reasoning by leveraging complementary strengths of diverse pre-trained models. Experiments on GSM8K, MathQA, and StrategyQA show consistent improvements over SOTA RLVR methods, with ablations confirming the contribution of each component.

References

Amini, A.; Gabriel, S.; Lin, S.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2357–2367. Minneapolis, Minnesota: Association for Computational Linguistics.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Fatemi, M.; Rafiee, B.; Tang, M.; and Talamadupula, K. 2025. Concise reasoning via reinforcement learning. *arXiv preprint arXiv:2504.05185*.

Feng, G.; Zhang, B.; Gu, Y.; Ye, H.; He, D.; and Wang, L. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36: 70757–70798.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lai, Y.; Lee, S.; Chen, G.; Poddar, S.; Hu, M.; Pan, D. Z.; and Luo, P. 2025. Analogcoder: Analog circuit design via training-free code generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 379–387.

Li, D.; Cao, S.; Cao, C.; Li, X.; Tan, S.; Keutzer, K.; Xing, J.; Gonzalez, J. E.; and Stoica, I. 2025. S*: Test time scaling for code generation. *arXiv preprint arXiv:2502.14382*.

Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Lu, J.; Wang, Y.; Yang, Z.; Liu, X.; Mac Namee, B.; and Huang, C. 2024. PaDeLLM-NER: parallel decoding in large language models for named entity recognition. *Advances in Neural Information Processing Systems*, 37: 117853–117880.

Lu, J.; Yang, L.; Namee, B.; and Zhang, Y. 2022. A Rationale-Centric Framework for Human-in-the-loop Machine Learning. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6986–6996. Dublin, Ireland: Association for Computational Linguistics.

Lu, J.; Yu, H.; Wang, Y.; Ye, Y.; Tang, J.; Yang, Z.; Wu, B.; Liu, Q.; Feng, H.; Wang, H.; Liu, H.; and Huang, C. 2025a. A Bounding Box is Worth One Token - Interleaving Layout and Text in a Large Language Model for Document Understanding. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 7252–7273. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Lu, J.; Yu, H.; Xu, S.; Ran, S.; Tang, G.; Wang, S.; Shan, B.; Fu, T.; Feng, H.; Tang, J.; et al. 2025b. Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient llm/mllm reasoning. *arXiv preprint arXiv:2505.15154*.

Lu, J.; Zhao, R.; Mac Namee, B.; and Tan, F. 2023a. Punifiedner: A prompting-based unified ner system for diverse datasets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 13327–13335.

Lu, J.; Zhu, D.; Han, W.; Zhao, R.; Mac Namee, B.; and Tan, F. 2023b. What Makes Pre-trained Language Models Better Zero-shot Learners? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2288–2303. Toronto, Canada: Association for Computational Linguistics.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Muennighoff, N.; Yang, Z.; Shi, W.; Li, X. L.; Fei-Fei, L.; Hajishirzi, H.; Zettlemoyer, L.; Liang, P.; Candès, E.; and Hashimoto, T. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Niu, K.; Chen, Z.; Yu, H.; Chen, Y.; Fu, T.; Zhao, M.; Li, B.; and Xue, X. 2025a. CReFT-CAD: Boosting Orthographic Projection Reasoning for CAD via Reinforcement Fine-Tuning. *arXiv preprint arXiv:2506.00568*.

Niu, K.; Yu, H.; Chen, Z.; Zhao, M.; Fu, T.; Li, B.; and Xue, X. 2025b. From Intent to Execution: Multimodal Chain-of-Thought Reinforcement Learning for Precise CAD Code Generation. *arXiv preprint arXiv:2508.10118*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qwen. 2023. Introducing Qwen-7B: Open foundation and human-aligned models (of the state-of-the-arts).

Setlur, A.; Nagpal, C.; Fisch, A.; Geng, X.; Eisenstein, J.; Agarwal, R.; Agarwal, A.; Berant, J.; and Kumar, A. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

Shah, K.; Dikkala, N.; Wang, X.; and Panigrahy, R. 2024. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Su, X.; Xie, S.; Liu, G.; Xia, Y.; Luo, R.; Jin, P.; Ma, Z.; Wang, Y.; Wang, Z.; and Liu, Y. 2025. Trust region preference approximation: A simple and stable reinforcement learning algorithm for llm reasoning. *arXiv preprint arXiv:2504.04524*.

Team, Q. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face*.

Team, S. 2025. Seed-Thinking-v1.5. Accessed: 2025-07-19.

Tong, P.; Brown, E.; Wu, P.; Woo, S.; IYER, A. J. V.; Akula, S. C.; Yang, S.; Yang, J.; Middepogu, M.; Wang, Z.; et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y.; Luo, Z.; Wang, J.; Zhou, Z.; Chen, Y.; and Han, B. 2025a. Eliciting Causal Abilities in Large Language Models for Reasoning Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15212–15220.

Wang, Y.; Yang, Q.; Zeng, Z.; Ren, L.; Liu, L.; Peng, B.; Cheng, H.; He, X.; Wang, K.; Gao, J.; et al. 2025b. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wen, L.; Cai, Y.; Xiao, F.; He, X.; An, Q.; Duan, Z.; Du, Y.; Liu, J.; Tang, L.; Lv, X.; et al. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*.

Xia, S.; Li, X.; Liu, Y.; Wu, T.; and Liu, P. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27723–27730.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Ye, Y.; Huang, Z.; Xiao, Y.; Chern, E.; Xia, S.; and Liu, P. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Yu, H.; Lu, J.; Wang, Y.; Li, Y.; Wang, H.; Huang, C.; and Li, B. 2025. Eve: Towards end-to-end video subtitle extraction with vision-language models. *arXiv preprint arXiv:2503.04058*.

Yu, H.; Wang, X.; Li, B.; and Xue, X. 2023. Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11943–11952.

Yue, Y.; Chen, Z.; Lu, R.; Zhao, A.; Wang, Z.; Song, S.; and Huang, G. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Zhang, C.; Feng, Z.; Zhang, Z.; Qiang, J.; Xu, G.; and Li, Y. 2025. Is LLMs Hallucination Usable? LLM-based Negative Reasoning for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1031–1039.

Zhang, X.; Du, C.; Pang, T.; Liu, Q.; Gao, W.; and Lin, M. 2024. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37: 333–356.

Zhao, R.; Meterez, A.; Kakade, S.; Pehlevan, C.; Jelassi, S.; and Malach, E. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

Zhou, H.; Luo, T.; and Jiang, Z. 2025. Core-to-Global Reasoning for Compositional Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10770–10778.