Thinking in Many Modes: How Composite Reasoning Elevates Large Language Model Performance with Limited Data

Zishan Ahmad
PhiLabs, Quantiphi Inc
Bengaluru, India
zishan.ahmad@quantiphi.com

Saisubramaniam Gopalakrishnan
PhiLabs, Quantiphi Inc
Bengaluru, India
gopalakrishnan.saisubramaniam@quantiphi.com

Abstract

Large Language Models (LLMs), despite their remarkable capabilities, rely on singular, pre-dominant reasoning paradigms, hindering their performance on intricate problems that demand diverse cognitive strategies. To address this, we introduce Composite Reasoning (CR), a novel reasoning approach empowering LLMs to dynamically explore and combine multiple reasoning styles like deductive, inductive, and abductive for more nuanced problem-solving. Evaluated on scientific and medical question-answering benchmarks, our approach outperforms existing baselines like Chain-of-Thought (CoT) and also surpasses the accuracy of DeepSeek-R1 style reasoning (SR) capabilities, while demonstrating superior sample efficiency and adequate token usage. Notably, CR adaptively emphasizes domain-appropriate reasoning styles. It prioritizes abductive and deductive reasoning for medical question answering, but shifts to causal, deductive, and inductive methods for scientific reasoning. Our findings highlight that by cultivating internal reasoning style diversity, LLMs acquire more robust, adaptive, and efficient problem-solving abilities.

1 Introduction

The remarkable progress in Large Language Models (LLMs) has led to significant advancements in natural language understanding and generation, largely due to their enhanced reasoning capabilities. While traditional approaches often rely on a single dominant reasoning paradigm, we hypothesize that enabling LLMs to leverage a diverse repertoire of reasoning strategies—such as deductive, inductive, abductive, and decompositional reasoning—can lead to more robust and accurate performance, particularly on complex tasks.

Recent state-of-the-art LLMs often employ Chain-of-Thought (CoT) prompting Wei et al. [2022], which elicits a step-by-step reasoning process. While effective, this process frequently exhibits a predominantly deductive flow. Our work, however, builds on the understanding that real-world problems demand greater flexibility, potentially requiring models to dynamically synthesize different reasoning methods. Previous efforts have explored enhancing LLM reasoning through improved decoding Wang et al. [2022], incorporating external knowledge Liu et al. [2023], or using techniques like self-consistency Naik et al. [2023]. However, there has been limited exploration of explicitly encouraging models to internally explore and integrate multiple distinct reasoning strategies during problem-solving.

To address this gap, we propose a novel composite reasoning (CR) approach. Our method encourages LLMs to explore and combine different reasoning strategies, allowing the model to consider multiple perspectives and pathways for more accurate and well-supported answers. We evaluate this approach

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

USER:
Question: All of the following are true about pulsus parvus EXCEPT Options:
[A] Seen in aortic stenosis
[B] Decreased left ventricular output
[C] Decreased peripheral resistance
[D] Decreased pulse pressure

AGENT:
<thought> Let's approach this question using a combination of deductive and abductive reasoning. Firstly, let's define what pulsus parvus is. Pulsus parvus refers to a weak and thready pulse that is often associated with cardiac tamponade, constrictive pericarditis, and hypovolemia. The key characteristic of pulsus parvus is a reduced amplitude of the pulse wave, which means the pulse is weak and difficult to feel. Now, let's consider each option to determine which one does NOT describe pulsus parvus: 1. **[A] Increased central venous pressure**: This can be true for conditions like cardiac tamponade, where fluid accumulates around the heart, increasing venous pressure and leading to a weak pulse. 2. **[B] Reduced stroke volume**: This is also consistent with conditions such as constrictive pericarditis, where the heart's ability to fill and eject blood is impaired, resulting in a weak pulse. 3. **[C] Decreased peripheral resistance**: This is the statement we need to critically evaluate. Peripheral resistance is related to how much the blood vessels oppose the flow of blood. A weak pulse is more indicative of a problem with the heart's pumping ability rather than the resistance of the peripheral vessels. 4. **[D] Low systolic blood pressure**: This is consistent with conditions that cause a weak pulse, such as hypovolemia or severe heart failure. By process of elimination and considering the nature of pulsus parvus, we can conclude that decreased peripheral resistance is not a characteristic of pulsus parvus. Pulsus parvus is primarily due to issues with cardiac output or filling pressures, not peripheral resistance. Thus, the correct answer is [C] Decreased peripheral resistance.
I started by defining pulsus parvus and then evaluated each option. Option C stands out because pulsus parvus is typically associated with issues in cardiac function (like reduced stroke volume or increased central venous pressure) rather than changes in peripheral resistance. I used deductive reasoning to eliminate options A, B, and D, leaving C as the outlier. </thought>
<answer> [C] Decreased peripheral resistance </answer>

Deductive Decompositional Causal Comparative Abductive

Figure 1: Example of Composite Reasoning style for a MedMCQA Question

using parameter-efficient fine-tuning (PEFT) techniques like LoRA Hu et al. [2022] and Group Relative Policy Optimization (GRPO) Shao et al. [2024] to assess performance in resource-constrained settings.

We conduct extensive experiments on three challenging and diverse datasets: MedMCQA Pal et al. [2022], MedXpertQA Zuo et al. [2025], and ARC-Complex Clark et al. [2018]. All fine-tuning and training were conducted using a maximum of 1,500 samples per dataset. We compare our CR strategy against standard CoT and a Standard Reasoning (SR) approach, demonstrating a compelling performance advantage. Furthermore, we show that GRPO with an outcome-based reward function (based solely on answer correctness) allows our CR approach to implicitly foster a more flexible and multi-faceted reasoning process that adapts to the specific demands of each domain.

We summarize our key contributions as follows: *(i).* A novel composite-reasoning approach that encourages LLMs to explore and adapt multiple reasoning strategies, *(ii).* We demonstrate the effectiveness of this approach on three challenging datasets within a resource-constrained training setting (maximum 1,500 samples), highlighting its superior sample efficiency, *(iii).* We show that GRPO with an outcome-based reward effectively guides our CR approach to explore and tailor diverse reasoning strategies to domain-specific needs, *and (iv).* Our results indicate significant performance improvements over standard CoT and SR baselines, highlighting the benefits of CR in terms of accuracy and token effectiveness in resource-constrained scenarios.

2 Methodology

This section details our experimental framework, which investigates the performance of our Composite Reasoning (CR) approach under resource constraints. All fine-tuning and training stages used a maximum of 1,500 samples from the official training splits of each dataset. We evaluate our models on the official test sets of ARC-Complex (1,119 questions), MedMCQA (4,183 questions) and MedXpertQA (950 questions).

We investigate three distinct reasoning paradigms for which initial trajectories were generated using a base Qwen-2.5-7B-Instruct model (except for SR, which was sourced from Deepseek-r1-7B Guo et al. [2025]).

2.1 Supervised Fine-Tuning (SFT) with LoRA

1. **Chain-of-Thought (CoT)** Wei et al. [2022]: This is the conventional method for eliciting sequential reasoning. The intuition is to prompt the model to "think step-by-step", which often leads to a logical, deductive-like progression that can improve accuracy on complex tasks.

Method	Reasoning	ARC-C		MedMCQA		MedXpertQA	
		Accuracy	Avg Token Length	Accuracy	Avg Token Length	Accuracy	Avg Token Length
Prompt	Direct	60.06%	38	45.88%	42	5%	35
	CoT	73.7%	254	52.77%	205	8%	287
	SR†	80.54%	515	48.96%	619	7.9%	1,139
	CR	83.10%	316	54.62%	335	7.8%	398
SFT	CoT	92.31%	252	55.35%	207	14.34%	282
	SR	75.20%	518	50.96%	612	11%	1,146
	CR	92.22%	320	55.20%	338	14.1%	405
GRPO	CoT	89.5%	249	55.10%	189	13.1%	412
	SR†	78.42%	524	49.23%	601	9.1%	1,110
	CR	90.35%	331	55.84%	331	10.08%	426
SFT + GRPO	CoT	93.85%	247	55.74%	208	14.63%	479
	SR	80.20%	518	51.80%	617	11.47%	1,112
	CR	**94.99%**	339	**56.30%**	313	15.9%	549

Table 1: Exact-Match Accuracy (%) on ARC-Complex, MedMCQA and MedXpertQA datasets across reasoning strategies and methods, with output token lengths. Best in **bold**. † experiments on *Deepseek-r1-7B*; others on *Qwen2.5-7B-Instruct*.

2. **Standard Reasoning (SR):** This baseline uses high-quality, pre-generated reasoning trajectories from Deepseek-r1-7B model. The purpose here is to test whether distilling a highly-polished reasoning style from a powerful external source is an effective fine-tuning strategy, even in a low-data setting.
3. **Composite Reasoning (CR):** Our approach explicitly prompts the model to dynamically explore and synthesize diverse reasoning strategies. The intuition behind this is to move beyond a single, linear thought process. It encourages the model to leverage a full "toolkit" of reasoning, including hypothesis generation (abduction), generalization (induction), and logical breakdowns (decomposition), thereby making it more adaptable to a wider range of problems. An illustrative example of a CR-generated thought process, showcasing these characteristics with annotated reasoning styles, is presented in Figure 1.

We finetune the base LLM using Supervised Fine-Tuning (SFT) using Low-Rank Adaptation (LoRA) on these generated trajectories. The intuition of SFT is to teach the model to imitate the reasoning styles we curated for each paradigm. This process essentially instills the desired "thinking patterns" (CoT, SR, or CR) into the model's behavior. Following SFT, we applied Group Relative Policy Optimization (GRPO) Shao et al. [2024], a reinforcement learning algorithm tailored for scenarios with sparse rewards. The core intuition of using GRPO with an outcome-based reward is to let the model self-refine its reasoning process based on a simple, yet powerful, signal: whether the final answer is correct or not. This encourages the model to generate reasoning that is not just plausible, but pragmatically effective at solving the task, without needing complex human-in-the-loop reward modeling for each reasoning step.

3 Results and Analysis

We present our empirical results on the ARC-Complex (ARC-C), MedMCQA, and MedXpertQA datasets. We analyze the performance of our proposed Composite Reasoning (CR) approach against Chain-of-Thought (CoT) and Standard Reasoning (SR) baselines. The detailed accuracy scores and average token lengths are presented in Table 1. In the direct zero-shot prompting setting, our Composite Reasoning (CR) prompt consistently outperforms standard Direct Prompting and CoT across all three datasets. While CR is slightly outperformed by SR on the highly complex MedXpertQA dataset (7.8% vs. 7.9%), this initial advantage on the other two datasets demonstrates the effectiveness of our prompt in eliciting stronger baseline reasoning. As indicated by the low overall accuracy scores, MedXpertQA is a significantly more difficult task requiring a higher level of domain-specific reasoning, which is reflected in the extremely verbose nature of the SR model's initial trajectories on this dataset (1,139 average tokens).

Supervised Fine-Tuning (SFT) using only 1,500 samples substantially enhances all strategies. Notably, CR SFT and CoT SFT achieve strong, competitive results, significantly outperforming SR SFT on all three datasets. The addition of GRPO to the SFT-tuned models consistently yields the highest performance. The CR SFT + GRPO configuration achieves the highest accuracy on both ARC-C (94.99%) and MedMCQA (56.30%), and it secures the top performance on MedXpertQA (15.9%).

Model	MedMCQA Acc. (%)
BioMistral-7B	40.2
OpenBioLLM-8B	54.1
ChatDoctor	31.5
PMC-LLaMA-7B	29.8
Baize-Healthcare	31.3
MedAlpaca-7B	32.9
Meditron-7B	31.1
PMC-LLaMA-13B	37.7
MedAlpaca-13B	35.7
ClinicalCamel	45.8
Huatuo-7B	24.8
HuatuoGPT-o1-8B	**60.4**
HuatuoGPT-o1-8B w/o RL	57.9
UltraMedical-8B	58.3
CR Prompt	54.62
CR SFT	55.20
CR GRPO	55.84
CR SFT + GRPO	**56.30**

Table 2: MedMCQA accuracy (%) comparison with existing baseline models from the 7-13B parameter size category. Note that CR is trained on only 1,500 training samples as compared to HuatuoGPT-o1 (40k samples) and UltraMedical (410k samples), yet remains competitive.

This demonstrates the potent synergy of CR with SFT and subsequent outcome-based reward tuning, enabling the model to explore and optimize reasoning paths better than CoT and SR to reach peak performance with limited training data. This synergy is particularly evident when analyzing the performance gain on the challenging MedXpertQA dataset. On this task, our CR method achieves a substantial gain of 8.1% (from 7.8% to 15.9%), which is significantly larger than the gains of CoT (6.63%) and SR (3.57%). This contrasts with the MedMCQA dataset, where the gain is more modest (1.68%), suggesting that CR's ability to learn from limited data is most pronounced when the problem requires deep, non-memorization-based reasoning.

Analysis of reasoning chain lengths reveals a compelling accuracy-verbosity trade-off. As noted, SR produces the longest reasoning paths, but this verbosity does not translate to higher accuracy in our fine-tuning setup. While CoT is generally the most concise, CR strikes a better balance, achieving superior accuracy with moderately longer but more effective reasoning chains. For the highly complex MedXpertQA dataset, the average token counts for both CR and CoT increase after GRPO training (from 405 to 549 for CR, and 282 to 479 for CoT), indicating that the models are generating more detailed reasoning to solve the harder problems. This suggests that GRPO optimizes CR towards more token-effective reasoning on simpler tasks while encouraging necessary verbosity for complex ones.

The sample efficiency of our method is particularly noteworthy on the MedMCQA dataset. As shown in Table 2, our CR SFT + GRPO model achieves an accuracy of 56.30%. This is highly competitive with medical LLMs like HuatuoGPT-o1-8B and UltraMedical-8B, despite our model being trained on only 1,500 samples, a small fraction of the 40k and 410k domain specific samples, respectively, used by those baselines. This highlights the remarkable sample efficiency of our CR approach. The analysis of reasoning style dynamics (visualized in Figures 2 and 3 in the Appendix) reveals that GRPO selectively modifies the problem-solving approaches of our models in a domain-dependent manner. For a more detailed discussion of these stylistic shifts, refer to Appendix A.3.

4 Conclusion

In this work, we introduce Composite Reasoning (CR), a method that enhances LLMs' complex reasoning by encouraging the exploration and integration of diverse strategies. In a resource-constrained (1,500-sample) LoRA-based fine-tuning setup on challenging datasets, including ARC-Complex, MedMCQA, and the highly demanding MedXpertQA, our CR approach consistently outperforms standard Chain-of-Thought (CoT) and Standard Reasoning (SR) baselines. This performance is particularly noteworthy on the most difficult tasks, where CR, combined with GRPO-based fine-tuning, achieves a significantly greater performance gain than other methods. Our experiments

underscore CR's remarkable sample efficiency, allowing it to compete with domain-specific LLMs on MedMCQA despite using orders of magnitude less training data. By encouraging diverse reasoning strategies like deductive, inductive, abductive, etc., our findings show that LLMs can develop more robust, adaptive, and effective problem-solving skills.

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. Diversity of thought improves reasoning abilities of large language models. *openreview*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Appendix

A.1 Experimental Setup

All experiments were conducted on a single NVIDIA A100 80GB GPU. All our experiments are based on Qwen-2.5-7B and Deepseek-r1-7B models both containing around 7 billion parameters. Each SFT training took around 7 hours on a single GPU, while GRPO tuning varied took between 24-48 hours for different experiments. We employed a consistent configuration across both the Supervised Fine-Tuning (SFT) and Generalized Reinforcement Preference Optimization (GRPO) phases. LoRA adapters were configured with a rank $r = 32$ and an alpha $\alpha = 64$, resulting in a scaling factor $s = \alpha/r = 2$. The target modules for LoRA integration included the following linear layers: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. These modules correspond to the query, key, value, and output projection layers in the attention blocks, as well as the feed-forward network components within the transformer architecture. During the SFT phase, we used a learning rate of 10^{-4}, a batch size of 8, and trained for 12 epochs. The optimizer employed was AdamW with a weight decay of 0.001, and the learning rate scheduler followed a linear warmup

with decay strategy. In the GRPO phase, a new LoRA adapter was trained using a learning rate of 10^{-4}, a batch size of 2, and for 1,500 training steps. The optimizer and scheduler mirrored those used in the SFT phase. This standardized set-up ensured consistency and comparability across our experimental evaluations. All the implementation was done in python utilizing unsloth, huggingface, trl and vllm packages.

A.2 Loss Functions and Optimization

The Supervised Fine-Tuning (SFT) process aims to minimize the standard auto-regressive language modeling loss (cross-entropy) over the reasoning trajectories. Let θ_{base} represent the frozen parameters of the base model and θ_{LoRA} represent the trainable LoRA adapter parameters. The SFT loss function is given by:

$$\mathcal{L}_{\text{SFT}}(\theta_{\text{LoRA}}) = -\sum_{j=1}^{|D_{\text{train}}|}\sum_{i=1}^{L_j}\log P(t_{j,i}|t_{j,1},\ldots,t_{j,i-1};\theta_{\text{base}},\theta_{\text{LoRA}}) \tag{1}$$

where D_{train} is the training set and L_j is the length of the trajectory T_j.

For the GRPO phase, a new LoRA adapter was trained with weights ϕ. The policy LLM is denoted as π_ϕ. For each input prompt x from the training dataset, the policy π_ϕ generates a group of M distinct reasoning trajectories ${\tau^{(m)}}_{m=1}^{M}$. A binary reward $R(\tau) \in 0, 1$ was assigned to each trajectory τ based on the exact match correctness of its final answer. The trajectory within the group of M generations that achieved the highest reward (i.e., a correct answer, if any) was identified as τ^*:

$$\tau^* = \underset{\tau \in {\tau^{(m)}}_{m=1}^{M}}{\text{argmax}} \; R(\tau) \tag{2}$$

GRPO updates the policy by comparing the rewards of trajectories within each group. The optimization objective is to maximize the expected relative reward, which encourages the model to favor trajectories with higher relative rewards without relying on an explicit value function.

A.3 Detailed Analysis of Reasoning Style Dynamics

The shift in reasoning style distribution due to GRPO, as depicted in the MedMCQA chart (Figure 3) compared to the ARC-C chart (Figure 2), underscores how simply using outcome-based optimization adapts reasoning strategies to domain-specific demands, often showing a stylistic alignment with human cognitive approaches.

On MedMCQA—a domain demanding diagnostic inference—Composite Reasoning with GRPO (CR-LoRA+GRPO) markedly amplifies Abductive reasoning (inferring best explanations) and Deductive reasoning (applying medical rules), making them the dominant styles. By contrast, on ARC-C, GRPO's CR primarily boosts Deductive and Causal reasoning, with only a modest uptick in Abductive and a stronger rise in Inductive reasoning. Chain-of-Thought post-GRPO on MedMCQA also increases Abductive and Deductive usage, but doesn't reach the peaks achieved by CR. Likewise, on ARC-C, GRPO steers CR toward Causal, Deductive, Decompositional, and Inductive reasoning—reflecting the general science emphasis on cause-effect, logical breakdown, and generalization—while Abductive reasoning remains less prominent than in the medical setting.

The Standard Reasoning (SR) strategy, on MedMCQA, much like on ARC-C, shows a less adaptive pattern post-GRPO, with several of its initially high general reasoning styles (like Causal and Comparative) potentially decreasing or not being effectively channeled into medically critical styles like Abductive reasoning.

This domain-dependent adaptation is synergistic with human expert reasoning. Physicians often employ a hypothetico-deductive process, generating hypotheses (abduction) and testing them against evidence and knowledge (deduction) Elstein et al. [1978]. The strong performance of the CR model and its post-GRPO reasoning profile in MedMCQA, with its emphasis on abductive and deductive styles, suggest that it learns to emulate these effective human diagnostic strategies more closely than other methods. Similarly, the broader scientific reasoning profile seen on ARC-C reflects the

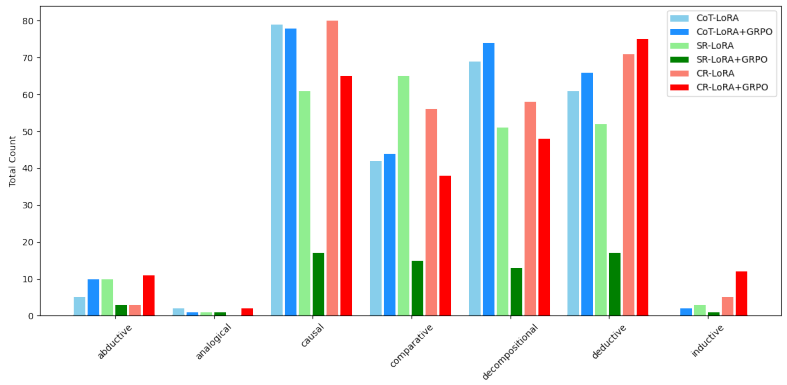

Figure 2: Analysis of reasoning strategies across the three trajectory types before and after applying GRPO tuning to the LoRA-tuned model on **ARC-C** dataset

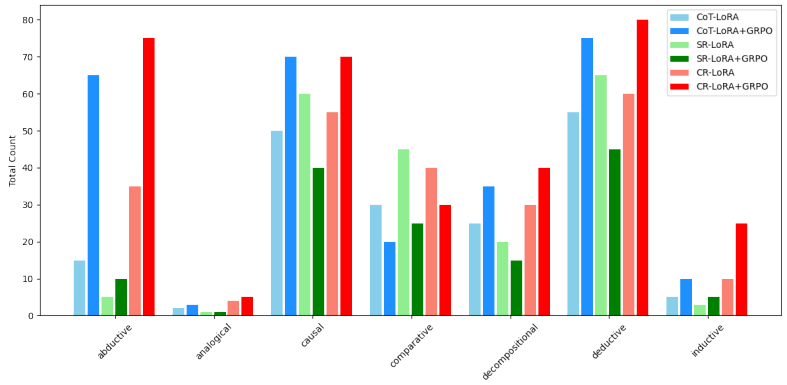

Figure 3: Analysis of reasoning strategies across the three trajectory types before and after applying GRPO tuning to the LoRA-tuned model on **MedMCQA** dataset

varied approaches humans use for general science problem-solving. The CR framework's flexibility, therefore, seems to allow GRPO to better identify and amplify the most effective, domain-appropriate human-like reasoning strategies.