Reasoning Patterns Matter: Enhancing Cross-Disciplinary Generalization in Large Language Models

Anonymous ACL submission

Abstract

While large language models (LLMs) have achieved strong results in mathematical reasoning, their ability to generalize reasoning skills across academic disciplines remains underexplored. This work investigates whether equipping models with diverse human reasoning patterns can improve cross-domain generalization.

We present RP-CD, a 2.28M instance dataset that annotates four reasoning patterns—verification, retrospection, decomposition, and reverse thinking—across diverse academic subjects. Based on RP-CD, we introduce Pattern Fine-Tuning (PFT), a method that injects explicit reasoning patterns into models enabling to internalize problem solving strategies.

Experiments with Qwen2.5-7B-Instruct demonstrate that PFT consistently outperforms supervised fine-tuning (SFT) across a range of benchmarks. On GSM8K, PFT-Mix improves accuracy by 5.7% over Vanilla-SFT, with larger gains observed on MATH (+4.6%) and BBH (+7.6%). PFT-Mix also demonstrates strong cross-lingual transfer on C-Eval (+5.3%) and improves instruction following performance on IFEval (+5.4%). Furthermore, after reinforcement learning (RL), the PFT-Mix-RL model achieves 35.5% accuracy on RP-CD, surpassing much larger models such as Qwen2.5-72B-Instruct (22.6%) and DeepSeek-R1-Distill-Qwen-32B (21.7%).

Our results highlight the value of reasoning patterns for enhancing cross-disciplinary and out-of-domain generalization. This suggests a practical path towards developing LLMs that are more robust in real-world, multi-domain applications.

1 Introduction

Recent advancements in reasoning within large language models (LLMs) (Jaech et al., 2024; DeepSeekAI et al., 2025; Team et al., 2025) have demonstrate significant progress, primarily in mathematics. However, their ability to generalize reasoning across diverse academic domains remains limited (Hochlehnert et al., 2025; Yue et al., 2025; AI et al., 2025). One possible reason is that current LLMs are not explicitly taught to use different types of reasoning (Gandhi et al., 2025). In contrast, humans solve problems across disciplines by flexibly employing diverse reasoning patterns.

This work delves into the question: *Can explicitly teaching models human-inspired reasoning patterns lead to more reliable generalization?* To explore this, we focus on four representative reasoning patterns commonly observed in human problem solving: verification (Koriat et al., 1980), retrospection (Gick and Holyoak, 1980), decomposition (Chi et al., 1981; Wang et al., 2023; Khot et al., 2023), and reverse thinking (Polya, 2014). As illustrated in Figure 1(a), these patterns form the backbone of our investigation into whether models can be taught not just to memorization (Chu et al., 2025), but to reason.

To investigate the impact of reasoning patterns, we introduce the Reasoning Patterns–Cross Disciplines (RP-CD) dataset, a large-scale benchmark comprising 2.28 M instances across STEM, social sciences, humanities, applied sciences, and other domains. Each question-answer pair is annotated with four distinct human-inspired reasoning patterns. We further introduce PFT, which trains models to answer questions using explicit reasoning patterns, rather than SFT relying solely on question–answer pairs.

Our experiments, summarized in Figure 1(c), reveal several key findings: (1) We observe clear synergistic gains from the diversity of reasoning patterns: on GSM8K (grade school math) (Cobbe et al., 2021b), PFT-Mix achieves a **5.7%** improvement, demonstrating the benefit of integrating multiple reasoning patterns; (2) On more challenging datasets, the advantages of PFT become more

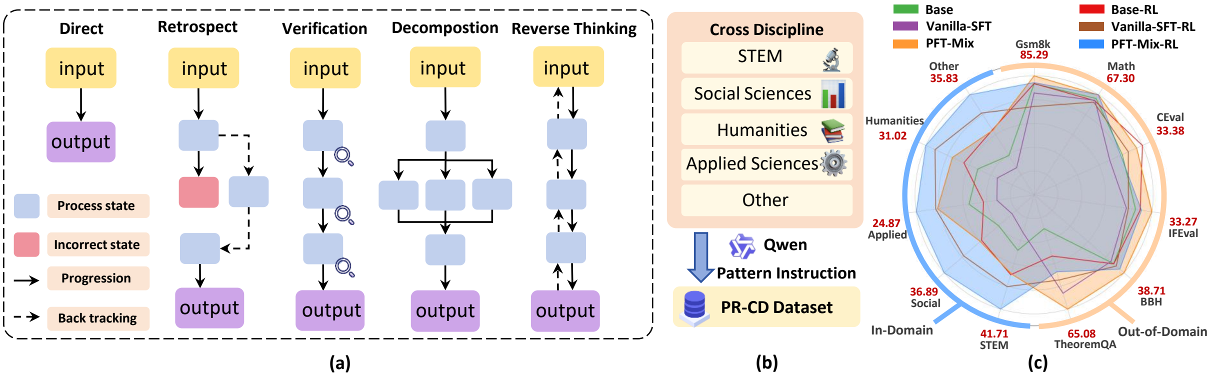

Figure 1: **Overview of reasoning patterns, dataset construction, and model performance.** **(a)** Four reasoning patterns: Retrospect (backtracking), Verification (step-by-step checking), Decomposition (divide-and-conquer), Reverse Thinking (goal-driven reasoning), and one direct output. **(b)** Reasoning trajectories are generated by Qwen2.5-72B-Instruct on the Multi-subject RLVR dataset to construct Reasoning Patterns – Cross Discipline Dataset (RP-CD). **(c)** PFT-Mix improves both in-domain and out-of-domain performance, outperforming Base and Vanilla-SFT. Base-Mix-RL surpasses Base-RL and Vanilla-SFT-RL on multiple metrics.

substantial—on MATH(competition mathematics) (Hendrycks et al., 2021a) and BBH(complex reasoning) (Suzgun et al., 2022), PFT-Mix delivers even larger improvements (**+4.6%** and **+7.6%**, respectively); (3) The benefits of PFT extend beyond English: we observe strong cross-lingual transfer on C-Eval (**+5.3%**) (Huang et al., 2023); (4) PFT also enhances instruction following and benefits subsequent reinforcement learning (RL) (Sutton and Barto, 2018), PFT-Mix-RL (7B) leading to a **35.5%** accuracy on RP-CD, substantially outperforming much larger models like Qwen2.5-72B-Instruct(Team, 2024)(**22.6%**) and DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025a)(**21.70%**).

2 Related Work

Reasoning Patterns in LLMs. Reasoning patterns have been shown to significantly enhance the reasoning capabilities of LLMs. (Wei et al., 2022b) first introduced the concept of Chain-of-Thought prompting, which guides models to emulate human-like reasoning processes. Building on this idea, methods such as Tree-of-Thought (Yao et al., 2023) and Graph-of-Thought(Besta et al., 2024) further improve performance by enabling models to explore diverse reasoning paths. The Self-Debate (Liang et al., 2024) framework advances this by encouraging models to engage in internal argumentation, thereby improving their ability to address open-ended questions. To generalize the benefits of such reasoning patterns across various domains, Self-Prompting was proposed, allowing models to dynamically select appropriate reasoning patterns based on the task at hand. (Gandhi et al., 2025) suggest that the presence of a reasoning pattern is more critical than its specific form. These approaches collectively demonstrate the effectiveness of reasoning patterns. In this work, we conduct a systematic study of such patterns and show that the learned reasoning patterns are not merely memorized templates (Allen-Zhu and Li, 2023; Ye et al., 2024; Kang et al., 2024), but rather represent intrinsic reasoning capabilities that generalize across domains.

Post-training. SFT (Zhang et al., 2022; Hoffmann et al., 2023; OpenAI, 2023; Google, 2023; Touvron et al., 2023) has become a standard approach for aligning LLMs to task-specific formats. FLAN (Wei et al., 2022a) demonstrated that instruction tuning enhances zero-shot generalization across diverse tasks, while LIMA (Zhou et al., 2024a) showed that SFT can also convey stylistic and structural preferences. More recent studies (Gandhi et al., 2024; Lehnert et al., 2024) incorporate reasoning traces—such as linearized solution paths or self-correction trajectories—into the SFT process to improve performance on mathematical and logical tasks (Ye et al., 2024; Qu et al., 2025; Kumar et al., 2024; Hwang et al., 2024; Han et al., 2024). RL (Ziegler et al., 2019; Ouyang et al., 2022; Sun et al., 2024; Ramamurthy et al., 2023; Abdulhai et al., 2023; Zhou et al., 2024b; Zhai et al., 2024) has also been explored for fine-tuning LLMs in domains like mathematical problem solving and code generation. (Luo et al., 2024)(Zeng et al., 2025; Lambert et al., 2025). However, these approaches often focus on narrow domains (Abdulhai et al., 2023; Zhou et al., 2024b; Zhai et al., 2024; Chen et al., 2024). In contrast, our work explores PFT, which demonstrates improved generalization and provide a more effective foundation for subsequent RL fine-tuning.

3 Dataset

3.1 Reasoning Pattern Definitions

To systematically investigate the role of reasoning patterns in learning, we define four core patterns, each grounded in cognitive science and educational theory, and observed in both human problem solving and model-generated reasoning. For each pattern, we design tailored prompt templates to elicit the desired reasoning process; detailed prompt examples are provided in Appendix A.

Verification involves deliberately stepwise checking and internal hypothesis testing. Inspired by dual-process theories of reasoning (Evans, 2003; Chi et al., 1981), this pattern encourages systematic evaluation at each step, reducing the likelihood of careless errors and fostering logical coherence throughout problem solving.

Retrospection means looking back and considering other possible solution paths, especially when the current approach fails. By learning from mistakes and being willing to try different methods (Gentner, 1983; Holyoak and Thagard, 1996; Chi et al., 1981), retrospection helps solvers adapt and find better ways to reach a solution.

Decomposition breaks complex problems into simpler, manageable sub-tasks, enabling a modular approach to inference. This pattern aligns with the classic divide-and-conquer principle in education (Polya, 2014; Sweller, 1988), which is widely recognized for reducing cognitive load and enhancing problem-solving efficiency in both humans and machines.

Reverse Thinking starts from a hypothesized goal or conclusion and works backward to uncover the necessary conditions or supporting premises. Closely related to backward reasoning and proof by contradiction (Evans et al., 2007; Polya, 2014), this strategy supports goal-driven, counterfactual, and non-linear reasoning, which are essential for creative problem solving and robust model inference.

Together, these four patterns capture a diverse array of cognitively plausible, human-inspired reasoning patterns, moving beyond the linear and monotonic generation typical of standard language models. Their inclusion provides a principled framework for studying how explicit reasoning pattern injection can enhance the learning and generalization of LLMs.

3.2 Dataset Construction

Based on our definition of four reasoning patterns, we construct **RP-CD**, a large-scale dataset for systematic study of reasoning patterns in LLMs. We start from the Multi-Subject-RLVR corpus (Su et al., 2025), a benchmark spanning multiple academic disciplines, which comprises 570k question–answer pairs with reference answers written by domain experts for examination purposes.

To create RP-CD,we leverage QWEN2.5-72B-INSTRUCT (Qwen et al., 2025) to generate four parallel reasoning patterns for each question–answer pair, covering Verification, Retrospection, Decomposition, and Reverse Thinking (see Appendix A). This process increases the dataset to 2.28 million instances, with every question associated with four distinct, pattern-specific solutions.

Crucially, RP-CD provides not only answers *what*, but also diverse, transparent demonstrations of *how* to solve each problem using different reasoning patterns. Through training on these varied problem-solving patterns, models can internalize more flexible reasoning habits, resulting in stronger out-of-domain generalization. RP-CD thus serves as a dedicated, large-scale resource for controlled study of reasoning pattern learning in LLMs (see Figure 1(a, b)). To foster reproducibility and future research, we will release the full dataset upon paper acceptance.

4 Experiment

At the heart of our study lies a central question: *How do reasoning patterns affect the learning and generalization of large language models*? To answer this, we design a set of experiments guided by five research questions, examining the effects of reasoning patterns across task difficulty, cross-domain and cross-lingual generalization, reinforcement learning, and comparisons with larger or distilled models. Together, these experiments build a comprehensive picture of reasoning pattern learning in LLMs.

4.1 Pattern Fine-Tuning

Vanilla SFT trains models to generate answers directly from questions, typically ignoring the reasoning patterns. PFT instead augments each training

Model	STEM	Social	Applied	Humanities	Others	Avg.
Qwen2.5-7B-Instruct (Base)	20.26	14.57	13.73	16.17	14.06	16.2
Vanilla-SFT	13.80	9.01	7.86	10.56	9.73	10.48
PFT-Rev	12.44	9.57	9.95	9.57	9.94	10.62
PFT-Veri	26.97	22.83	21.59	22.44	24.21	24.06
PFT-Decom	22.19	21.45	17.74	24.09	22.62	21.23
PFT-Retro	23.24	21.08	16.37	16.83	20.08	20.42
PFT-Mix	28.76	22.08	20.39	23.43	23.05	24.09
Qwen-72B-Instruct	20.10	28.7	20.50	21.00	25.20	22.60
DeepSeek-R1-Distill-Qwen-32B	23.20	21.80	26.70	20.50	18.50	21.70
Base-RL	29.17	21.51	10.67	20.13	23.25	21.90
Vanilla-SFT-RL	33.52	28.89	20.78	28.38	29.28	28.71
PFT-Mix-RL	**41.71**	**36.89**	**24.87**	**31.02**	**35.83**	**35.47**

Table 1: In-distribution performance on the RP-CD test set across five academic domains: STEM, Social Sciences (Social), Applied Sciences (Applied), Humanities, and Others. **Before RL**, PFT-Mix achieves an average score of **24.09%**, outperforming Base (**16.2%**) and Vanilla-SFT (**10.48%**), which is trained without reasoning supervision. Despite being a 7B model, PFT-Mix also surpasses larger models such as Qwen2.5-72B-Instruct (22.60%) and DeepSeek-R1-Distill-Qwen-32B (21.70%). **After RL**, PFT-Mix-RL further improves to **35.47%**, achieving the highest overall performance across all domains.

example with a human-inspired reasoning pattern. Each instance is represented as a triplet (Q, P_i, A), where Q is the question, P_i is a reasoning pattern chosen from verification, decomposition, retrospection, or reverse thinking, and A is the answer.

The model is trained to maximize the probability of generating the combined reasoning pattern and answer sequence $y = [P_i; A]$ given the question Q:

$$\mathcal{L}_{\text{PFT}} = -\mathbb{E}_{(Q,P_i,A)\sim\mathcal{D}} \sum_{j=1}^{t} \log p(y_j \mid Q, y_{<j})$$

Here, $p(y_j \mid Q, y_{<j})$ is the probability assigned to the j-th token in the output, and t is the total length of the reasoning pattern plus answer.

The key idea of PFT is not simply to expose the model to more data, but to repeatedly present reasoning patterns in the learning process so that the model internalizes these patterns and learns to apply them in pursuit of the final answer. Through this repeated exposure, PFT encourages the model to move beyond superficial pattern recognition or answer memorization, fostering genuine acquisition and flexible use of reasoning patterns as part of its problem solving process.

4.2 Group Relative Policy Optimization

We follow DeepSeek-R1 (DeepSeek-AI et al., 2025b) and employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our reinforcement learning algorithm. GRPO applies policy gradients, calculated from the reward loss, to optimize model parameters.

Our reward function consists of two components: a **Format Reward** for outputs that strictly follow the required structure (reasoning enclosed in <think>...</think> and answers in <answer>...</answer>), and a **Correctness Reward** assigned when the answer is accurate, as judged by Qwen2.5-72B-Instruct (Team, 2024).

$$R = \begin{cases} 1, & \text{if answer is correct} \\ 0.1, & \text{if only format is correct} \\ 0, & \text{otherwise} \end{cases}$$

All reinforcement learning experiments in this work use GRPO, including the Base-RL, Vanilla-SFT-RL, and PFT-RL configurations. This unified setup allows for a fair comparison of different training setups under the same RL framework.

4.3 Implementation Details

All experiments are conducted using Qwen2.5-7B-Instruct (Yang et al., 2024) as the base model, with the entire pipeline implemented via LLaMA-Factory (Zheng et al., 2024) and Verl (Sheng et al., 2024) to ensure reproducibility and scalability. Training was performed on 8×A100 GPUs, totaling 5,760 GPU hours across all configurations (see Appendix 3 for details). We consider

Model	GSM8K	MATH	C-Eval	IFEval	BBH	TheoremQA
Qwen2.5-7B-Instruct (Base)	79.61	64.78	55.94	28.65	57.10	9.88
Vanilla-SFT	73.01	63.58	48.02	26.06	55.27	28.75
PFT-Retro	74.37	67.14	65.46	31.05	62.02	27.25
PFT-Rev	60.58	67.30	63.48	22.00	55.14	**36.88**
PFT-Decom	76.95	66.80	64.55	30.68	**67.18**	35.50
PFT-Veri	40.71	66.16	68.44	29.21	64.54	23.00
PFT-Mix	**85.29**	67.24	65.08	**33.27**	64.70	33.38
Base-RL	79.76	63.32	**68.71**	29.94	57.27	17.88
Vanilla-SFT-RL	63.31	62.34	59.78	26.62	59.08	24.88
PFT-Mix-RL	80.44	**67.31**	54.97	30.31	61.64	22.50

Table 2: Evaluation of out-of-domain (OOD) generalization across six diverse benchmarks, covering arithmetic reasoning (GSM8K, MATH), chinese multi-discipline (C-Eval), instruction-following (IFEval), complex multi-step reasoning (BBH), and theorem-driven question answering (TheoremQA).

several model variants: the base model; Vanilla-SFT, trained solely on question–answer pairs from RP-CD, without any explicit reasoning pattern injection, and four PFT variants, each corresponding to a distinct human-inspired reasoning pattern—verification (PFT-Veri), decomposition (PFT-Decom), retrospection (PFT-Retro), and reverse thinking (PFT-Rev). To further explore the diversity of reasoning patterns, we also include PFT-Mix, which randomly samples instances from all four patterns, exposing the model to a broad spectrum of reasoning styles. All models are evaluated on both in-domain (RP-CD) and six out-of-domain benchmarks covering arithmetic (Cobbe et al., 2021a), mathematics (Hendrycks et al., 2021a),Chinese multi-subject QA (Huang et al., 2023), instruction following (Zhou et al., 2023), complex reasoning (Suzgun et al., 2022), and theorem-driven question answering (Chen et al., 2023) (see Appendix D for details).All evaluations are run on the OpenCompass (Contributors, 2023) platform to ensure consistency and reproducibility. We next present and analyze model performance on these benchmarks (Table 1, Table 2).

4.4 Quantitative Insights

Models trained with diverse reasoning patterns consistently outperform baselines(base model, vanilla SFT), as shown in both Table 1 (in-domain) and Table 2 (out-of-domain). On RP-CD, PFT-Mix achieves a substantial gain over the base, with reinforcement learning (PFT-Mix-RL) further elevating performance to **35.47%**, outperforming substantially larger models. Out-of-domain, the improvements are robust across all benchmarks, with the largest gains observed on tasks requiring multi-hop reasoning (BBH) or theorem-driven question answering (TheoremQA). These results highlight that explicit supervision with varied reasoning patterns enables models to develop transferable skills, enhancing generalization well beyond simple answer memorization.

5 Insights

We present our empirical findings structured around five research questions (RQ1–RQ5), each designed to examine a different aspect of reasoning pattern effectiveness. For each RQ, we report experimental results on relevant benchmarks and provide corresponding analysis. All experiments use Qwen2.5-7B-Instruct (Team, 2024) as the base model.

5.1 RQ1: Does the Diversity of Reasoning Patterns Matter?

To investigate whether training on diverse reasoning patterns benefits generalization. As show in Figure 2, we evaluate four single-pattern PFT variants—Verification, Retrospection, Decomposition, and Reverse Thinking—on GSM8K (Cobbe et al., 2021c). As shown in Figure 2, all individual variants underperform the base model (**79.61%**), with PFT-Retrospection (**74.37%**) and PFT-Decomposition (**76.95%**) exceeding SFT (**73.01%**), while PFT-Verification (**40.71%**) and PFT-Reverse (**60.58%**) lag behind.

In contrast, PFT-Mix which jointly trains on all four reasoning styles, achieves **85.29%**, outperforming the base model by **5.68%**. This striking result suggests a synergistic effect: instead

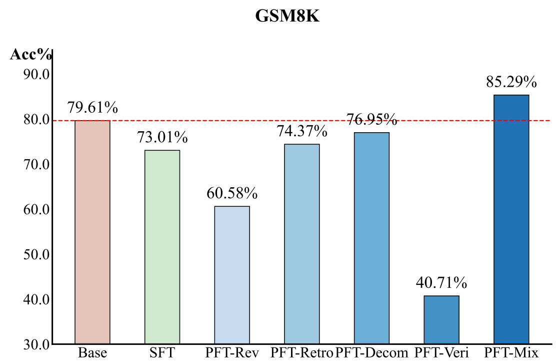

Figure 2: **Performance on GSM8K.** The red dashed line marks the base model's performance. While all single-pattern PFT variants underperform the base, PFT-Mix—trained on a diverse set of reasoning patterns—achieves the highest accuracy **(85.29%)**. This result suggests that reasoning pattern diversity leads to a non-trivial synergy that enhances model generalization.

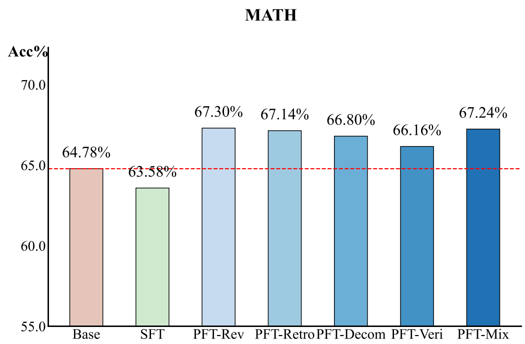

Figure 3: **Performance on MATH.** As task complexity increases, SFT begins to lag behind the base model, indicating its limited generalization. In contrast, all PFT variants exhibit consistent improvements, with PFT-Rev, PFT-Retro, and PFT-Mix achieving the highest accuracies. This suggests that reasoning patterns are more robust to out-of-domain even in moderately difficult settings.

of merely averaging out the strengths and weaknesses of each strategy, their combination yields a performance gain beyond any individual pattern. If no interaction effect existed, we would expect PFT-Mix to fall between the strongest and weakest variants—yet it surpasses both. This highlights the importance of cognitive diversity in enhancing model generalization.

5.2 RQ2: Are Reasoning Patterns More Effective for Harder Tasks?

We hypothesize that reasoning patterns are not only beneficial in general, but become increasingly essential as task complexity rises. On the MATH (Hendrycks et al., 2021b) benchmark, which contains more formal and symbolic problems than GSM8K (Cobbe et al., 2021c), all PFT-trained models show improved performance compared to the base model, while vanilla SFT exhibits a decline. This suggests that structured reasoning patterns help models generalize beyond surface-level memorization, even when the task difficulty begins to increase. Notably, PFT-Rev, PFT-Retro, and PFT-Mix outperform both the base and SFT models by a clear margin (see in Figure 3).

We further evaluate performance on BBH (Suzgun et al., 2022), which features even more complex multi-hop reasoning and abstract inference. Here, the gap between SFT and PFT becomes even wider: PFT-Decomposition leads with a **10.08%** absolute gain, while most other PFT variants improve by more than **5%**(Figure 4).The growing performance gap across harder tasks provides empirical support for our hypothesis: the utility of reasoning patterns scales with task complexity.

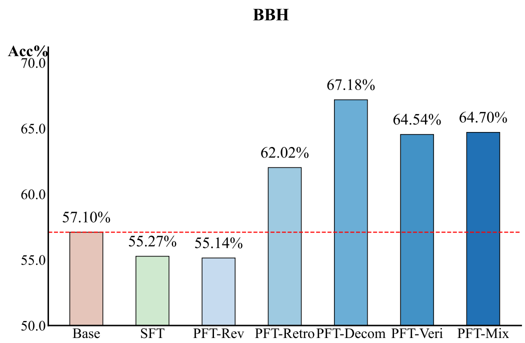

Figure 4: **Performance on BBH.** On the even more challenging BBH benchmark, the benefits of reasoning patterns become more pronounced: nearly all PFT variants surpass the base by a wide margin, with PFT-Decom leading by **+10.08%**. The gap between SFT and PFT widens further, underscoring the increasing value of pattern-driven reasoning as task difficulty and required generalization grow.

5.3 RQ3: Are Reasoning Patterns Transferable Across Tasks and Languages?

To evaluate whether reasoning patterns support the generalization and flexible application of knowledge, we assess model performance on two unseen benchmarks: TheoremQA (Chen et al., 2023), a multi-domain dataset containing formal scientific questions grounded in expert-written theorems across mathematics, physics, computer science, and finance; and C-Eval (Huang et al., 2023), a

Chinese benchmark spanning 52 disciplines. These benchmarks allow us to examine transferability both across task formats and linguistic boundaries.

As shown in Figure 5, reasoning patterns significantly enhance model performance on TheoremQA. Since this benchmark requires integrating domain knowledge from multiple scientific disciplines to answer theorem-driven questions, improvements here signal that models are not merely fitting seen tasks, but have acquired reusable reasoning capabilities that generalize beyond the training distribution. Notably, PFT-Rev and PFT-Decom achieve the strongest results, suggesting that step-reversal and problem decomposition are particularly effective patterns for applying knowledge in novel problem settings.

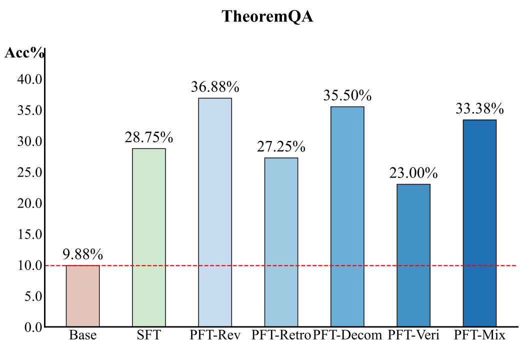

Figure 5: **Performance on TheoremQA.** All PFT variants outperform Base and vanilla SFT, with PFT-Rev achieving the highest accuracy (**36.88%**). This indicates that models trained with reasoning patterns are better equipped to apply their learned knowledge in unfamiliar scientific domains.

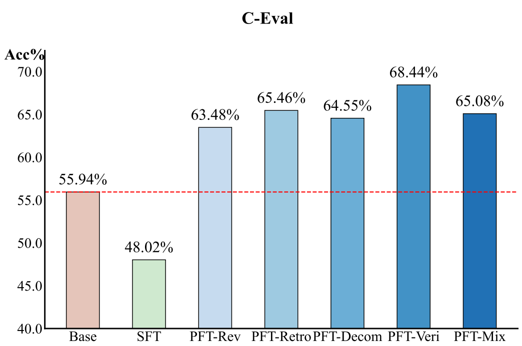

Figure 6: **Performance on CEval.** Although trained solely on English data, all PFT variants generalize to Chinese QA. PFT-Veri achieves the best performance (**68.44%**), demonstrating that reasoning pattern supervision transfers across languages.

Figure 6 further demonstrates the robustness of reasoning patterns in a cross-lingual setting. Despite being trained exclusively on English data, all PFT models significantly outperform the Base and SFT baselines on the Chinese multi-subject C-Eval benchmark. This suggests that the benefit of reasoning patterns is not language-specific, but instead stems from their ability to guide problem-solving in a generalizable way. Among the variants, PFT-Veri—which emphasizes self-checking at each reasoning step—achieves the best results, underscoring the utility of verification-based reasoning in diverse linguistic and disciplinary contexts.

5.4 RQ4: Do Reasoning Patterns Improve Reinforcement Learning?

Prior studies suggest that SFT improves instruction-following ability, which in turn benefits reinforcement learning (RL) (Chu et al., 2025). However, we find that vanilla SFT fails to provide a strong initialization for RL. As shown in Figure 7, PFT-trained models, especially PFT-Mix, demonstrate superior performance on IFEval both before and after RL training.

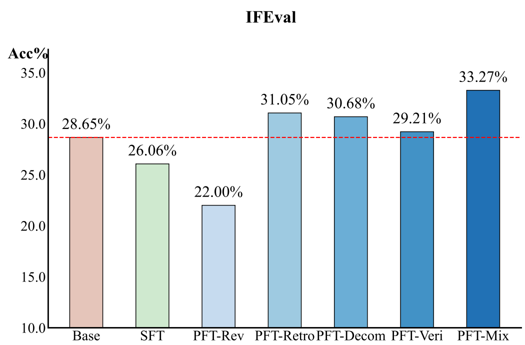

Figure 7: **Performance on IFEval.** A benchmark for instruction following. Most PFT variants outperform both the base and SFT models, with PFT-Mix achieving the best result (**33.27%**). This confirms that reasoning pattern supervision enhances instruction-following ability—an essential precursor for downstream reinforcement learning.

To verify this, we compare three configurations: base-RL, vanilla SFT-RL, and PFT-Mix-RL. The PFT-Mix model achieves the highest in-distribution accuracy after RL (**35.37%**), outperforming vanilla SFT-RL (**28.71%**) and base-RL (**21.90%**). This shows that reasoning pattern pretraining enables models to better leverage reward signals during RL, likely due to better-structured intermediate reasoning.

5.5 RQ5: Can PFT-Trained Small Models Outperform Larger or Distilled Models?

Beyond improving reinforcement learning, we further investigate whether our 7B model trained with reasoning patterns can outperform much larger or distilled models. As shown in Figure 8, PFT-Mix-RL (7B) achieves the highest accuracy (**35.47%**) on RP-CD, surpassing Qwen2.5-72B-Instruct (**22.60%**) and DeepSeek-R1-Distill-Qwen-32B (**21.70%**), despite their larger sizes and access to long COT supervision.

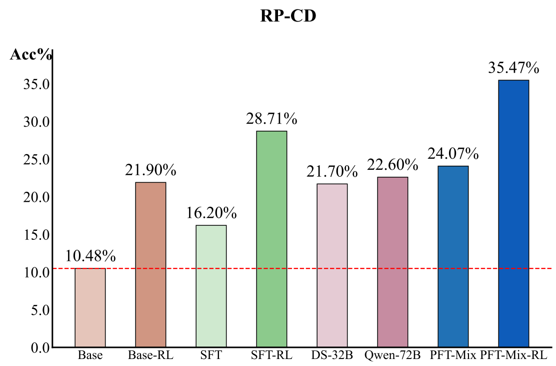

Figure 8: **Performance on RP-CD dataset.** PFT-Mix-RL achieves the highest accuracy (**35.47%**), outperforming not only its own SFT counterpart (**24.07%**) but also much larger models such as Qwen2.5-72B (**22.60%**) and DeepSeek-R1-Distill-Qwen-32B (**21.70%**). Notably, Base, SFT, and PFT each gain **11%** from RL, highlighting consistent post-RL improvement. These results support RQ4 (PFT boosts RL learning) and RQ5 (smaller models with PFT can surpass larger or distilled models).

These results underscore the effectiveness of reasoning patterns: even with sequences capped at 512 tokens, our models outperform those trained on thousands of tokens. This suggests that performance gains stem not from scale or context length, but from the quality of reasoning supervision. Compact yet cognitively meaningful signals prove sufficient to drive strong generalization. Furthermore, we observe a consistent **11%** accuracy boost from RL across Base, SFT, and PFT variants, confirming that PFT enhances a model's ability to benefit from RL.

Summary of Results. Our experimental results provide consistent evidence that incorporating human-inspired reasoning patterns into training substantially enhances model performance across tasks and domains. The benefits of PFT,particularly PFT-Mix,extend well beyond in-domain gains, enabling stronger generalization to new domains, languages, and out-of-domain reasoning challenges. Notably, small models trained with reasoning patterns can match or even outperform much larger or distilled models on RP-CD. These findings underscore the central role of reasoning patterns in model training and suggest that incorporating reasoning pattern offers an effective and accessible path toward more robust and generalizable language models.

6 Conclusion

We present the first large-scale, systematic study of *how training large language models with human-inspired reasoning patterns impacts learning and generalization*. By introducing the RP-CD dataset and PFT approach, we show that equipping models with diverse reasoning patterns leads to substantial improvements in both in-domain and out-of-domain settings, compared to Vanilla-SFT.

Concretely, our PFT trained Qwen2.5-7B-Instruct (Base model) achieves **35.5%** accuracy on the in-domain RP-CD benchmark, outperforming much larger models such as Qwen2.5-72B-Instruct (**22.6%**) and DeepSeek-R1-Distill-Qwen-32B (**21.7%**). On challenging out-of-domain tasks, PFT brings absolute gains of **+7.6%** on BBH and **+27%** on TheoremQA over standard baselines.

These results demonstrate that reasoning pattern learning is a general and transferable mechanism for improving model generalization. To enable LLMs to perform real-world, multidisciplinary reasoning, learning human-inspired reasoning patterns is essential. Our findings underscore the central message of this work: **Reasoning Pattern Matters**.

Limitations

While our study demonstrates the effectiveness of reasoning pattern supervision in enhancing both in-domain performance and cross-domain generalization, several limitations remain.

First, our proposed RP-CD dataset provides explicit reasoning trajectories with a fixed maximum length, and we do not systematically examine how sequence length influences the quality or complexity of reasoning. It is plausible that longer reasoning paths may enable models to express more elaborate thought processes, but this remains an open question for future exploration.

Second, the PFT-Mix training configuration increases cognitive diversity by sampling different reasoning patterns across examples. However, each example in the dataset still follows a single reasoning style. In reality, humans often switch between multiple reasoning patterns within a single problem. Our work does not investigate such intra-instance hybrid reasoning, which could further enhance model flexibility.

Third, while our evaluation covers a diverse set of out-of-domain (OOD) tasks, we report performance primarily at the aggregate level. In-depth analyses—such as per-task breakdowns for BBH or per-discipline breakdowns for C-Eval may uncover more nuanced effects of each reasoning pattern. We leave this fine-grained analysis to future work.

Finally, our study focuses on four reasoning patterns—verification, retrospection, decomposition, and reverse thinking—chosen for their cognitive plausibility and broad applicability. Yet we acknowledge that human reasoning is richer and more domain-specific. Disciplines such as law, medicine, or philosophy may involve distinct patterns (e.g., analogical reasoning, diagnostic inference) that are not covered in our work. A comprehensive taxonomy of reasoning patterns remains an important direction for extending this research.

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. LMRL Gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*.

Essential AI, :, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, and 10 others. 2025. Rethinking reflection in pre-training. *Preprint*, arXiv:2504.04022.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Jie Chen, Xintian Han, Yu Ma, Xun Zhou, and Liang Xiang. 2024. Unlock the correlation between supervised fine-tuning and reinforcement learning in training code large language models. *arXiv preprint arXiv:2406.10305*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. *Preprint*, arXiv:2305.12524.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021c. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeekAI and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Jonathan St BT Evans, Simon J Handley, Helen Neilens, and David E Over. 2007. Thinking about conditionals: A study of individual differences. *Memory & cognition*, 35(7):1772–1784.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *Preprint*, arXiv:2503.01307.

Kanishk Gandhi, Denise HJ Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah Goodman. 2024. Stream of search (sos): Learning to search in language. In *First Conference on Language Modeling*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Mary L. Gick and Keith J. Holyoak. 1980. Analogical problem solving. *Cognitive Psychology*, 12(3):306–355.

DeepMind Google. 2023. Introducing Gemini: Our largest and most capable AI model.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2024. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. *Preprint*, arXiv:2411.14794.

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

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. *Preprint*, arXiv:2103.03874.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. 2025. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *Preprint*, arXiv:2504.07086.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2023. Training compute-optimal large language models. *NeurIPS*.

Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. 2024. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. *arXiv preprint arXiv:2404.10346*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Katie Kang, Amrith Setlur, Dibya Ghosh, Jacob Steinhardt, Claire Tomlin, Sergey Levine, and Aviral Kumar. 2024. What do learning dynamics reveal about generalization in LLM reasoning? *arXiv preprint arXiv:2411.07681*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Lucas Lehnert, Sainbayar Sukhbaatar, DiJia Su, Qinqing Zheng, Paul Mcvay, Michael Rabbat, and Yuandong Tian. 2024. Beyond a*: Better planning with transformers via search dynamics bootstrapping. *arXiv preprint arXiv:2402.14083*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate.

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2024. Wizardcoder: Empowering code large language models with evol-instruct. In *ICLR*.

OpenAI. 2023. GPT-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

George Polya. 2014. How to solve it: A new aspect of mathematical method. In *How to solve it*. Princeton university press.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. 2025. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

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.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. *arXiv preprint arXiv:2407.20311*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Preprint*, arXiv:2504.13837.

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.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2024a. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

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

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024b. ArCHer: Training language model agents via hierarchical multi-turn RL. *arXiv preprint arXiv:2402.19446*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompt Templates for Data Construction.

To explicitly teach models human-inspired reasoning strategies, we design structured prompt templates for four core reasoning patterns: **Verification**, **Reverse Thinking**, **Retrospect**, and **Decomposition**. These patterns reflect cognitively grounded problem-solving strategies and guide the model to reason in interpretable and systematic ways. Each prompt wraps intermediate reasoning in <think> tags and final answers in <answer> tags, promoting consistency, traceability, and structured output generation.

- **Verification** (Figure 9): Encourages step-by-step reasoning with explicit validation at each stage, reducing logical errors and enhancing reliability.
- **Reverse Thinking** (Figure 10): Starts from the answer and defines backward subgoals to reach the solution, fostering goal-driven deduction.
- **Retrospect** (Figure 11): Uses backward chaining to identify necessary preconditions, helping the model explore alternative reasoning paths.
- **Decomposition** (Figure 12): Breaks problems into subcomponents and uses trial-and-error to test and refine solution paths.

These templates are used throughout the construction of the RP-CD dataset and during PFT, enabling the model to internalize diverse and generalizable reasoning strategies across domains.

B Score Prompt

To automatically evaluate the correctness of model-generated answers, we design **Score Prompt** as show in Figure 13. This prompt compares the final output of a model's reasoning process against a provided reference answer and produces a binary decision: `YES` if the answers match, or `NO` otherwise. To ensure consistency and prevent hallucinated justifications, the prompt explicitly restricts outputs

Table 3: **Hyperparameter for RL and SFT.**

Hyperparameter	RL	SFT
train_batch_size	1024	128
ppo_mini_batch_size	256	–
ppo_micro_batch_size_per_gpu	16	–
n_samples_per_prompt	4	–
max_epochs	1	1
prompt_max_len	512	–
generate_max_len	1024	–
max_len	–	4096
actor_learning_rate	1e-6	–
init_kl_coef	0.001	–

Table 4: **PR-CD Dataset Analysis.**

	Avg.	Med.	Std.
Retrospection	247.5	224.0	81.48
Verification	360.3	357.0	97.0
Decomposition	248.1	233.0	80.2
Reverse Thinking	140.4	130.0	62.92
Mix	249.1	228.0	112.6
Base	5.6	4.0	5.17

to be only YES or NO, with no explanation. This evaluation mechanism supports automatic large-scale scoring of generated reasoning traces across multilingual and multi-domain datasets.

C Training Details

The training details, including key hyperparameters used during the PFT and reinforcement learning (RL) stages, are summarized in Table 3.

D Benchmarks Description

In-domain: RP-CD. The RP-CD (Reasoning Patterns – Cross Discipline) test set contains 6,000 multi-disciplinary questions covering STEM, Social Sciences, Applied Sciences, Humanities, and Others. Each question is annotated with a reference answer and four types of reasoning patterns, enabling a systematic evaluation of model generalization within the training distribution.

Out-of-domain:

- **GSM8K** (Cobbe et al., 2021a): Consists of 8,500 high-quality, grade-school level math word problems requiring multi-step arithmetic reasoning. Each question is paired with a detailed solution rationale and answer.

- **MATH** (Hendrycks et al., 2021a): A large-scale benchmark of 12,500 problems drawn from high school and competition mathematics, spanning algebra, calculus, geometry, probability, and more. Each problem is accompanied by a detailed step-by-step solution.

- **C-Eval** (Huang et al., 2023): A Chinese multi-subject exam benchmark containing 13,948 multiple-choice questions across 52 disciplines, including STEM, social sciences, humanities, and others. C-Eval assesses factual knowledge and subject-specific reasoning in Chinese.

- **IFEval** (Zhou et al., 2023): An instruction-following evaluation set with 500 open-ended prompts designed to test models' ability to follow instructions and provide verifiable, controllable outputs.

- **BBH** (Big-Bench Hard) (Suzgun et al., 2022): Comprises 23 challenging tasks with a total of 4,435 questions covering diverse skills, such as logic, mathematics, code generation, causal reasoning, and abstract language understanding. Many tasks require multi-hop or compositional reasoning.

- **TheoremQA** (Chen et al., 2023): Contains 800 theorem-driven question-answering problems grounded in 350+ real-world theorems from mathematics, physics, chemistry, and other domains. Each question is constructed to require understanding and application of a specific theorem, testing both factual recall and deductive reasoning.

E PR-CD Dataset Analysis

Table 4 shows the statistics of token counts across different reasoning patterns on the PR-CD dataset. The table provides the average (Avg.), median (Med.), and standard deviation (Std.) of token counts for each model variant.

Table 5 provides a demonstration of the PR-CD dataset with examples from different domains, such as Social Sciences, STEM, Humanities, and Applied Sciences. Each entry includes the question type, the question itself, and the corresponding answer, offering a broad spectrum of academic disciplines and reasoning challenges.

Table 5: **PR-CD Dataset Demonstration.**

Type	question	answer
Social Sciences	The aesthetician who describes aesthetic experience as 'the activity of appreciation' is	Schopenhauer
STEM	An industrial enterprise consumes 5 million tons of fresh water annually, reuses 1.1 million tons of process water, recycles 500,000 tons of cooling water, and reuses 1 million tons of wastewater. The process water intake is 1.5 million tons, and the supplementary water for the indirect cooling water system is 700,000 tons. Therefore, the amount of industrial water reused by the enterprise is ______.	2.6 million tons
Humanities	The supervisors and other directly responsible personnel who are directly responsible for the illegal acts of forging or altering accounting vouchers, in addition to administrative penalties, ( ).	if they are state personnel, should also be given administrative sanctions
Applied Sciences	The quality of food protein is mainly determined by	Protein content, nitrogen balance, protein digestibility, biological value of protein, net protein utilization

Table 6: Performance on Out-Domain C-EVAL Datasets by Model

Model	**CEval-Stem**	**Social**	**Humanities**	**Other**	**Hard**	**Avg.**
Qwen2.5-7B-Instruct (Base)	49.82	72.59	58.47	49.40	53.17	55.94
Vanilla-SFT	45.77	62.31	46.81	40.34	51.17	48.02
PFT-Rev	58.28	74.30	66.67	59.91	58.33	63.48
PFT-Veri	**64.16**	81.88	71.31	61.15	**62.60**	68.44
PFT-Decom	59.83	75.46	68.53	59.22	60.06	64.55
PFT-Retro	58.93	81.39	71.65	56.68	56.62	65.46
PFT-Mix	61.23	77.22	68.01	58.10	56.41	65.08
Base-RL	58.62	**83.37**	**75.32**	**67.14**	55.40	**68.71**
Vanilla-SFT-RL	59.03	65.73	59.96	55.54	55.47	59.78
PFT-Mix-RL	46.46	69.91	56.76	55.07	46.38	54.97

Discipline Distribution in RP-CD Test Set. Figure 22 shows the distribution of academic disciplines in the RP-CD test set, which contains 6,000 instances sampled for evaluation purposes. While this distribution does not reflect the full RP-CD dataset, it covers a wide range of subject areas, including Basic Medicine (9.9

F Performance on Out-Domain CEval.

As shown in Table 6, the performance of various models on the CEVL out-of-domain datasets is presented. Among the models, PFT-Veri achieves the highest performance in several domains, including CEval-Stem and Social with an average score of 68.44. PFT-Mix shows strong results across all domains, with an average of 65.08. The PFT-Mix-RL model, which combines reasoning patterns with reinforcement learning, performs particularly well in the Social domain, achieving 69.91%. These results highlight the effectiveness of reasoning patterns and RL in enhancing model generalization across diverse domains.

G Performance of PFT-Mix-RL Using Different Reasoning Patterns on RP-CD Test Set

We present the performance of the PFT-Mix-RL mqodel using different reasoning patterns on the RP-CD test set. As shown in Figures 14, 15, 16, and 17, the figures demonstrate the effectiveness of each reasoning pattern (Verification, Decomposition, Retrospection, Reverse Thinking).

H Performance of PFT-Mix Using Different Reasoning Patterns on RP-CD Test Set

These figures show the performance of the PFT-Mix model. As shown in Figures 18, 19, 20, and 21, the figures present the performance with each reasoning pattern.

I Ethical Considerations

Our study aims to understand whether human-inspired reasoning patterns can help language models generalize more effectively across domains. Although the work does not directly engage high-risk applications or safety-critical systems, we believe that it is important to reflect on the broader implications of releasing reasoning-augmented datasets and models.

In building RP-CD, we chose to focus on four cognitive patterns—verification, retrospection, decomposition, and reverse thinking—that are frequently used in human problem-solving. These reasoning trajectories were not hand-annotated, but generated through prompt-based interactions with Qwen-72B-Instruct. This decision allowed us to scale data creation, but it also means that the explanations are synthetic in nature. We reviewed samples throughout the process to avoid egregious errors or inappropriate content, but the dataset inevitably inherits some of the imperfections of the underlying model. Readers should be mindful that these trajectories—while structured and seemingly deliberate—do not reflect human reasoning in any rigorous cognitive sense. They are heuristics, not ground truth.

One ethical tension we encountered is the apparent "validity" that these explanations might suggest. When a model presents its reasoning in clear, natural language, it becomes easier for users to overestimate the reliability or intentionality of its outputs. This is especially relevant when reasoning appears consistent but is actually flawed—a known issue in contemporary LLMs. We include structured outputs not to imply correctness, but to study how training on such scaffolding may shape generalization. The risk of misplaced trust, however, remains, especially if this work were to be adapted without appropriate caution.

Another limitation arises from the linguistic and epistemic scope of our dataset. Although the underlying questions cover a wide range of academic disciplines, all the reasoning trajectories are in English and generated by a single model family. This may introduce systemic biases, not only in linguistic style but also in the kinds of reasoning patterns the model tends to prefer. We have not yet explored fairness across disciplines or whether certain domains receive more coherent or credible reasoning support than others. These are meaningful open questions that we hope future work will address.

We release RP-CD and the associated models to support research on interpretable and cognitively grounded NLP. However, we do not advocate direct deployment in real-world decision-making settings, particularly in domains like healthcare, law, or education, without thorough safety evaluations and human oversight. Our goal is not to present a deployable system, but to offer a resource for better understanding how models can be taught to reason and what the limitations of that reasoning may be.

All resources will be released under a noncommercial research license to support transparency and further academic exploration. We emphasize that the dataset and models are intended solely for research purposes.

Verification Prompt

You are a multi-domain expert assistant.
Your task is to generate a step-by-step reasoning trajectory that leads to the provided answer using the **Verification** reasoning pattern.

Instructions:

- You are given a question and its correct answer.
- Use Verification reasoning: at each step, explicitly verify whether your intermediate result satisfies the problem's criteria.
- Ensure each <think> step is logically valid, independently verifiable, and builds toward the solution.
- Wrap each step in a <think> tag.
- Wrap the final answer in an <answer> tag.

Output Format:

<think> Step-by-step reasoning with explicit verification at each step. </think>
<answer> write answer here </answer>

Here is the problem, answer:
Question: {question}
Answer: {answer}

Figure 9: **Demonstration for Verification Prompt**

Reverse Thinking Prompt

You are a multi-domain expert assistant.
Your task is to generate a step-by-step reasoning trajectory that leads to the provided answer using the **Subgoal Setting** reasoning pattern.

Instructions:

- You are given a question and its correct answer.
- Use Subgoal reasoning: clearly define intermediate subgoals and solve them step-by-step to reach the final goal.
- Each <think> block must specify a subgoal and solve it explicitly.
- Wrap each step in a <think> tag.
- Wrap the final answer in an <answer> tag.

Output Format:

<think> Subgoal 1: define and solve the first meaningful subgoal. </think>
<think> Subgoal 2: define and solve the next subgoal. </think>
<answer> write answer here </answer>

Here is the problem, answer:
Question: {question}
Answer: {answer}

Figure 10: **Demonstration for Reverse Thinking Prompt.**

Retrospect Prompt

You are a multi-domain expert assistant.
Your task is to generate a step-by-step reasoning trajectory that leads to the provided answer using the **Backward Chaining** reasoning pattern.

Instructions:

- You are given a question and its correct answer.
- Use Backward Chaining: start from the final goal (answer) and work backward, identifying and solving necessary preconditions.
- Each <think> step must state what condition must be true to reach the next step, moving backward toward known facts.
- Wrap each backward step in a <think> tag.
- Wrap the final answer in an <answer> tag.

Output Format:

<think> Goal: state the final answer. What must be true immediately before this step? </think>
<think> Step backward: identify the previous condition or value needed. </think>
<answer> write answer here

Here is the problem, answer:
Question: {question}
Answer: {answer}

Figure 11: **Demonstration for Retrospect Prompt.**

Decompostion Prompt

You are a multi-domain expert assistant.
Your task is to generate a step-by-step reasoning trajectory that leads to the provided answer using the **Backtracking** reasoning pattern.

Instructions:

- You are given a question and its correct answer.
- Use Backtracking reasoning: try plausible steps, and if they fail, explicitly discard them and try alternative paths.
- Clearly mark which paths are abandoned and why.
- Wrap each attempt in a <think> tag, including both successful and failed paths.
- Wrap the final answer in an <answer> tag.

Output Format:

<think> Attempt 1: try a plausible step. If it fails, explain and discard. </think>
<think> Attempt 2: try an alternative path. Justify why this works. </think>
<answer> write answer here </answer>

Here is the problem, answer:
Question: {question}
Answer: {answer}

Figure 12: **Demonstration for Decomposition Prompt.**

Score Prompt

Given a problem, determine whether the final answer in the provided (incomplete) solution process matches the reference answer. The reference answer may be one single option character (e.g., A, B, C, D), a numerical value, an expression, or a list of answers if multiple questions are involved. **The reference answer may be in Chinese or another language ,but your evaluation should be language-agnostic.**

Your task:

- Compare the final output of the solution process with the reference answer.
- If they **match exactly** , output **YES**.
- If they **do not match**, output **NO**.
- If the solution process is unclear, incomplete, or ambiguous, assume it is incorrect and output **NO**.

Your output must be strictly **YES** or **NO**, with no additional words, punctuation , or explanation.

Question: {question}
Solution Process (Final Step Only): {prediction}
Reference Answer: {answer}
Output:

Figure 13: **Demonstration for Score Prompt.**

Verification GRPO thinking

Question: The product process design activities carried out by machinery manufacturing enterprises belong to ______.

Think: To solve this using the Verification thinking method, we begin by forming a **hypothesis** about the category to which product process design activities belong, and then **verify** it against known definitions and classifications.

Step 1: Formulate the hypothesis

Product process design in machinery manufacturing likely falls under a broader category of **production or manufacturing activities** — but more specifically, it likely belongs to **production preparation** or **technological preparation**.

Step 2: Verify definition of product process design

Product process design refers to the planning and detailing of how a product will be manufactured, including:
Selection of equipment
Process steps
Tooling
Production flow
Quality control methods
This is typically done **before** mass production begins.

Step 3: Verify standard industrial classification

In manufacturing enterprises, these activities are generally classified under **production technical preparation**, also known as **manufacturing process planning or engineering**.
This category includes:
Process route design
Fixture and tooling planning
Work instruction preparation
All these are consistent with the role of **product process design**.

Step 4: Confirm

Therefore, based on standard manufacturing practices and definitions, we can **verify** that product process design activities belong to **production technical preparation**.

Answer: Production technology preparation
GT: Production technology preparation process

Figure 14: **The performance of PFT-Mix-RL using verification on the RP-CD test set.**

Decomposition GRPO thinking

Question: In the introduction stage of the product life cycle, the strategy of launching new products with high prices and high promotional costs belongs to ( ).

Think: To solve this problem using **Decomposition thinking**, we break the problem into smaller, manageable parts:

Understand the Product Life Cycle (PLC):

The PLC includes four main stages:

- Introduction
- Growth
- Maturity
- Decline

Focus on the Introduction Stage:

In the **introduction stage**, the product is newly launched. Key features of this stage:

- High costs (due to R&D, marketing, etc.)
- Low sales volume
- Need for heavy promotion to build awareness
- Uncertain customer acceptance

Interpret the Strategy in Question:

The strategy involves:

- **High prices**
- **High promotional costs**

Identify Common Marketing Strategies in the Introduction Stage:

There are generally two major strategies:

- **Skimming Pricing Strategy:** High price, high promotion — aimed at early adopters to recover development costs quickly.
- **Penetration Pricing Strategy:** Low price, high promotion — aimed at gaining market share quickly.

Match the Description:

High prices and high promotional costs" matches the **Skimming Pricing Strategy**, not penetration.

Answer: Skimming Pricing Strategy

GT: Skimming Pricing Strategy

Figure 15: **The performance of PFT-Mix-RL using decomposition on the RP-CD test set.**

Retrospect GRPO thinking

Question: An industrial enterprise consumes 5 million tons of fresh water annually, reuses 1.1 million tons of process water, recycles 500,000 tons of cooling water, and reuses 1 million tons of wastewater. The process water intake is 1.5 million tons, and the supplementary water for the indirect cooling water system is 700,000 tons. Therefore, the amount of industrial water reused by the enterprise is ______.

Think: To solve this problem using **Retrospect** thinking, we look backward from the goal — finding the **amount of industrial water reused** — and determine what contributes to that reuse. We consider all reused water sources and eliminate overlaps or misinterpretations.

Step 1: Understand the components

Fresh water consumed: 5 million tons/year
Process water reused: 1.1 million tons/year
Cooling water recycled: 0.5 million tons/year
Wastewater reused: 1 million tons/year
Process water intake: 1.5 million tons/year
Supplementary cooling water: 0.7 million tons/year

Step 2: Define "industrial water reused"

This term generally includes:
Reused process water
Recycled cooling water
Reused wastewater
These are all forms of water that have already been used once and are **used again** in the industrial process.

Step 3: Add up reused water sources

Reused process water: **1.1 million tons**
Recycled cooling water: **0.5 million tons**
Reused wastewater: **1.0 million tons**

Step 4: Total industrial water reused

Total = 1.1 + 0.5 + 1.0 = **2.6 million tons**

Answer: 2.6 million tons
GT: 2.6 million tons

Figure 16: **The performance of PFT-Mix-RL using retrospect on the RP-CD test set.**

Reverse GRPO thinking

Question: The regulation that only the answers and works of the five great jurists on legal issues have legal effect is______

Think: To apply **Reverse Thinking**, we consider the problem from the opposite perspective. Instead of directly identifying the name of the regulation that gives legal effect only to the works of five great jurists, we ask:
What kind of legal system **does not** give legal effect to all jurists' opinions?
Under what circumstances would **only a select few** jurists' writings be considered legally binding?
Which legal tradition historically elevated the writings of a few jurists above others?
Thinking backwards:
If the writings of **only** five jurists are considered legally binding, then this implies a **restrictive** and **authoritative** selection within a legal framework.
This matches the concept in **Roman Law**, particularly during the time of **Emperor Theodosius II** and **Valentinian III**, who issued the **Law of Citations** in 426 AD. This law gave the writings of **five jurists**—Gaius, Papinian, Paulus, Ulpian, and Modestinus—**legal authority** in court.
Therefore, the regulation in question is the **Law of Citations**.

Answer: Law of Citations
GT: Citation Law

Figure 17: **The performance of PFT-Mix-RL using reverse thinking on the RP-CD test set.**

Verification SFT thinking

Question: What is regarded as the 'last frontier' for modern enterprises to reduce costs? ( )

Think: To solve this using the **Verification** thinking method, we begin by forming an initial **hypothesis** and then **verify** it through established knowledge in business management and cost control practices.
Step 1: Formulate Hypothesis
In modern enterprises, most traditional cost-reduction strategies—like material cost control, labor efficiency, and production optimization—have been extensively implemented. However, in recent years, **logistics** has gained attention as an area where significant cost savings are still possible.
Thus, a reasonable hypothesis is: **logistics** is considered the "last frontier" for cost reduction.
Step 2: Verify with Business Literature
Business and supply chain management literature often refers to logistics, particularly **supply chain and logistics optimization**, as having untapped potential for:
Reducing transportation and warehousing costs
Improving delivery efficiency
Minimizing inventory holding costs
Enhancing customer satisfaction and responsiveness
These benefits have a direct impact on overall enterprise cost structure.
Step 3: Confirm Wording
The phrase **"last frontier"** implies the final major area left for cost-cutting, after other avenues have been explored. Multiple business sources and industry experts indeed describe **logistics** as the last great opportunity for modern companies to reduce operational costs.
Step 4: Conclusion
Verification confirms that among all business operations, **logistics** is widely regarded as the "last frontier" for cost reduction in modern enterprises.

Answer: logistics
GT: logistics

Figure 18: **The performance of PFT-Mix using verification on the RP-CD test set.**

Decomposition SFT thinking

Question: Male, 50 years old, complete edentulous, moderate alveolar ridge height, scattered bone spines in the maxilla, slightly enlarged maxillary tuberosity on the left side, normal interarch distance. The treatment to be done before restoration is ______

Think: To solve this problem using Decomposition thinking, we need to break the case down into smaller, manageable parts. Let's analyze each component of the case:

Patient Demographics:

Male, **50 years old** – This tells us the age and gender but has limited direct impact on the pre-prosthetic treatment unless systemic conditions are present.

Oral Condition:

Completely edentulous – Indicates the absence of all teeth, requiring a full-arch prosthesis (e.g., complete dentures or implant-supported prosthetics).

Moderate alveolar ridge height – This suggests there is enough bone for denture support but not ideal. The ridge may need preservation, reshaping, or enhancement for stability and retention.

Scattered bone spines in the maxilla – Bone spicules or sharp ridges can cause discomfort and may interfere with denture seating. These often require **alveoloplasty** (surgical smoothing).

Slightly enlarged maxillary tuberosity (left) – An enlarged tuberosity can interfere with denture placement and occlusion. If excessive, it may need **tuberosity reduction**.

Normal interarch distance – This is favorable for prosthesis fabrication; no need for vertical dimension correction.

Treatment Goal:

Treatment **before restoration** implies **pre-prosthetic surgery**, i.e., procedures to prepare the oral cavity for future prosthetic rehabilitation.

Now, let's deduce the necessary pre-prosthetic treatments:

Alveoloplasty to smooth scattered bone spines in the maxilla.

Maxillary tuberosity reduction (left side) if the enlargement interferes with prosthesis fit.

Evaluation of ridge form – While moderate height is acceptable, ensure it's smooth and broad enough for denture support.

Answer: The treatment to be done before restoration includes:

Alveoloplasty to remove or smooth the scattered bone spines in the maxilla.

Surgical reduction of the left maxillary tuberosity if it impedes denture placement or occlusion.

General pre-prosthetic ridge contouring to optimize the foundation for a complete denture.

No intervention is needed for interarch space, but thorough evaluation and planning remain essential.

GT: Remove bone spines

Figure 19: **The performance of PFT-Mix using decomposition on the RP-CD test set.**

Retrospect SFT thinking

Question: The key to sickle cell anemia is the mutation of the gene, where the amino acid changes from glutamic acid to?

Think: To solve this problem using **Retrospect** thinking, we start from the **known result**—sickle cell anemia—and trace back to the **specific molecular change** that causes it.
Step 1: What is sickle cell anemia?
Sickle cell anemia is a genetic blood disorder caused by abnormal hemoglobin, specifically **hemoglobin S**.
Step 2: What causes hemoglobin S?
Hemoglobin S results from a **point mutation** in the gene that encodes the beta chain of hemoglobin.
Step 3: What is the specific mutation?
In sickle cell anemia, there is a **single nucleotide substitution** in the DNA:
The codon GAG (which codes for **glutamic acid**) is changed to GTG.
GTG codes for **valine**.
Step 4: Amino acid change
So, the amino acid in the sixth position of the beta-globin chain changes from **glutamic acid** (hydrophilic) to **valine** (hydrophobic), causing the hemoglobin molecules to stick together and distort red blood cells into a sickle shape.

Answer: valine
GT: valine

Figure 20: **The performance of PFT-Mix using retrospect on the RP-CD test set.**

Reverse SFT thinking

Question: The layering degree of the mortar should be about cm.

Think: To apply **Reverse Thinking**, we start by flipping the problem:
Instead of asking *"What should the layering degree (thickness) of the mortar be?"*, we ask:
"What happens if the layering degree is not appropriate?"
If the mortar layer is **too thin**, it may not bond the bricks properly, leading to weak structural integrity.
If the mortar layer is **too thick**, it may shrink excessively while drying, leading to cracks and instability.
So now we ask:
"What is the range that avoids both extremes?"
By thinking from consequences backward, we aim to find the optimal balance. Standard construction guidelines suggest that a mortar joint (layering) is usually around **1 cm**, as it provides both bonding strength and stability without excessive shrinkage.
Thus, by reverse thinking through the *undesirable outcomes*, we deduce the correct layering thickness.

Answer: 1 cm
GT: 1-2 cm

Figure 21: **The performance of PFT-Mix using reverse thinking on the RP-CD test set.**

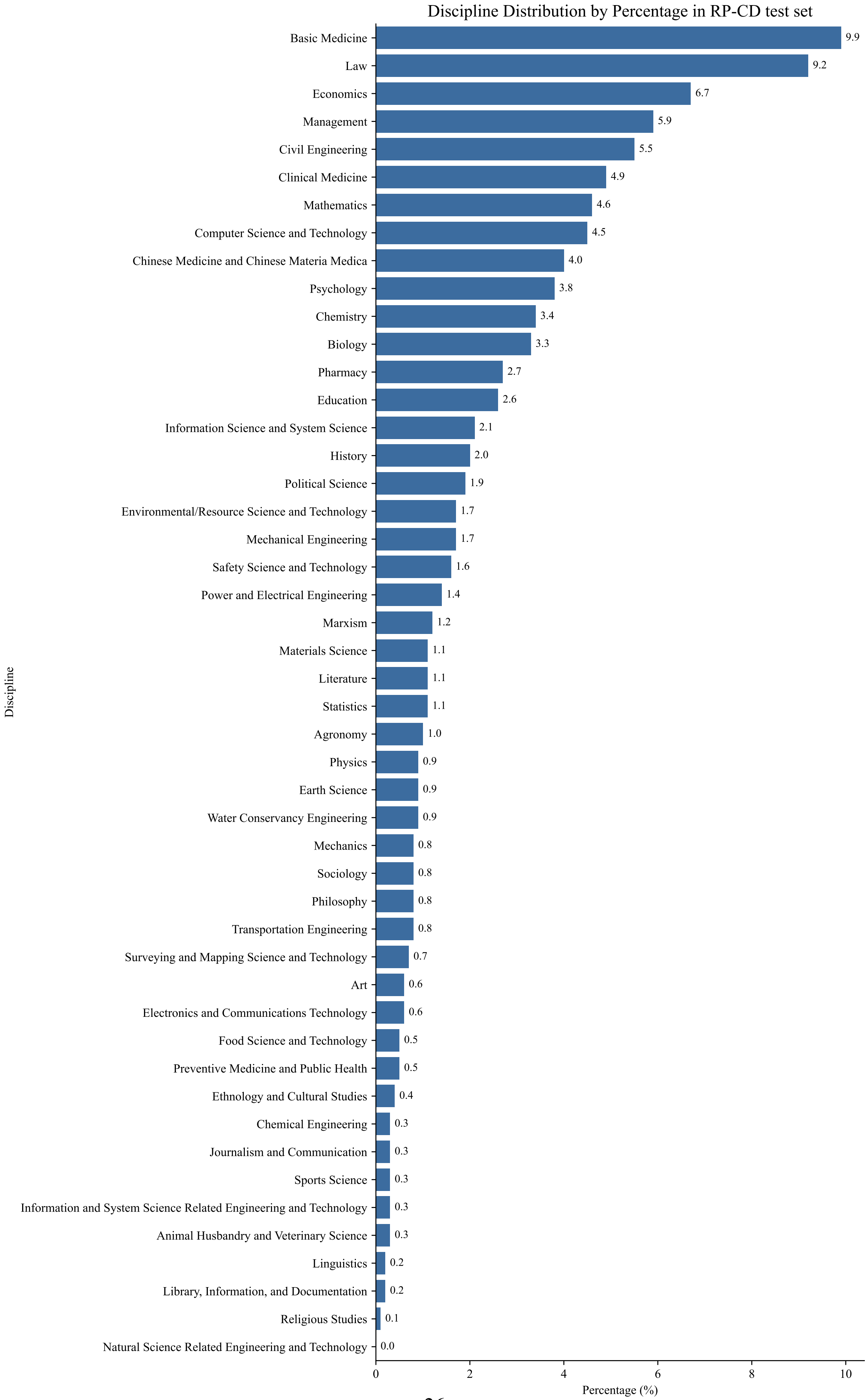

Figure 22: **Distribution of subject occurrences in the test set of RP-CD**