PARADIGMCOT: PARADIGM-GUIDED CHAIN-OF-THOUGHT IN LARGE LANGUAGE MODELS VIA POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-Thought (CoT) prompting may exhibit reasoning-path drift, logical inconsistency, and fabricated explanations, limiting the reliability of LLM reasoning. While structured prompting and process supervision improve local correctness and factuality, they provide limited holistic, principle-level guidance for the overall reasoning trajectory. Drawing on established instructive paradigms in both general and domain-specific fields, we propose **ParadigmCoT**: a general and flexible framework that injects instructive paradigms into the CoT process via policy optimization. At its core is an automated paradigm localization and evaluation module, which identifies paradigm-triggering steps and evaluates step-wise adherence throughout the reasoning chain. These paradigm alignment signals, combined with factuality and answer correctness, are incorporated into a fine-grained advantage adjustment in reinforcement learning to promote paradigm-aligned reasoning. ParadigmCoT is instantiated with general and medical paradigms, and evaluate it across diverse benchmarks and three LLM backbones. It achieves substantial gains over base models (up to 40.2% on general tasks and 34.5% on medical tasks), and outperforms RL-based baselines by 2.5%–3.4% and 5.8%–7.7% on average, respectively. This framework offers a new perspective on improving the quality and reliability of LLM-generated CoT, contributing to more reliable AI.

1 INTRODUCTION

Large language models (LLMs), empowered by Chain-of-Thought (CoT) prompting techniques, have significantly advanced complex reasoning by encouraging stepwise decomposition (Wei et al. (2022); Chen et al. (2023); Liu et al. (2023)). Nevertheless, conventional CoT mainly makes intermediate reasoning explicit, but offers little principled guidance for coordinating individual steps across the full inference trajectory. As a result, LLM-generated reasoning chains may drift away from relevant evidence, problem constraints, or appropriate reasoning strategies, leading to logically inconsistent transitions, unsupported post hoc rationalizations, or factually correct but irrelevant statements (Xu et al. (2024); Huang et al. (2024); Lindsey et al. (2025)), as shown in Fig. 1. These errors are particularly problematic in long-horizon, multi-step reasoning settings that require sustained evidence tracking and constraint satisfaction, and in high-stakes domains such as medicine and law, where even small reasoning errors can have substantial consequences. These failures severely undermine the credibility of LLMs in real-world applications. Therefore, it is urgent to develop a general approach that steers the overall reasoning trajectory toward more principled, coherent, and reliable inference.

To improve the quality of CoT reasoning in LLMs, recent research has shifted from outcome-based to process-based supervision (Lightman et al. (2023); Wang et al. (2024); Tarek & Beheshti (2025); Uesato et al. (2022)). Approaches such as process reward models and step-wise reinforcement learning (e.g., Math-Shepherd , FSPO) provide more fine-grained signals—leveraging atomic statement accuracy and final answer correctness—to effectively densify feedback and mitigate reward exploitation (Wang et al. (2024); Li & Ng; Farquhar et al.). In parallel, structured prompting frameworks such as Tree of Thoughts (ToT) (Yao et al. (2023)) and Graph of Thoughts (GoT) (Besta et al. (2024)) explicitly organize the reasoning process: ToT expands and evaluates multiple reasoning paths in a tree, while GoT generalizes this to arbitrarily structured graphs, facilitating feedback loops. These methods typically grow tokens with breadth or depth to carry intermediate state, which has effectively

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

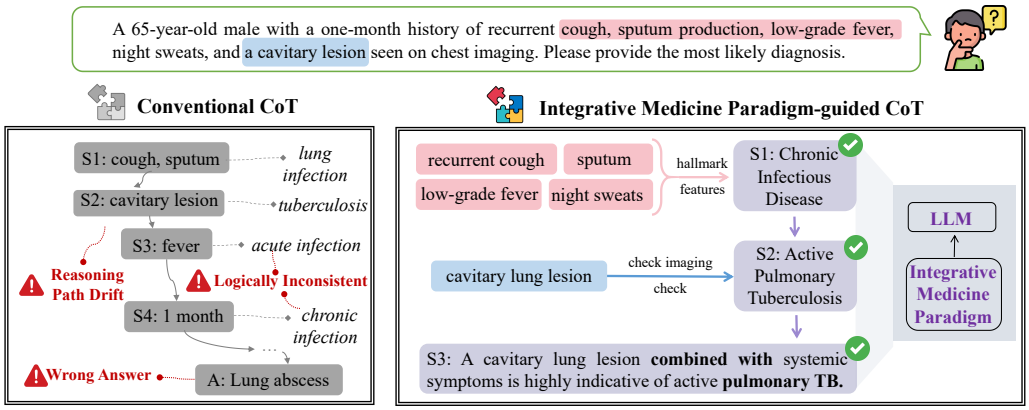


Figure 1: Conventional CoT shows reasoning drift and inconsistency, resulting in unreliable answer. Paradigm-CoT uses instructive paradigms to guide the whole CoT process, achieving holistic, principled reasoning.

inspired the incorporation of symbolic or structural scaffolds. Despite these benefits, they mainly focus on local correctness or structural exploration, while lacking global paradigm-level guidance for maintaining coherence across the reasoning trajectory of LLMs.

Over time, *human society has developed diverse paradigms—both general and domain-specific—that support systematic reasoning and strategic problem-solving* (McArthur (1972); Hastie (1984); Pearl (2022)). Drawing on these paradigms, we propose **ParadigmCoT**, a general and flexible framework that **integrates instructive paradigms into LLM-generated CoT, steering the overall trajectory toward more principled reasoning**, as shown in the lower-right of Fig. 1. We instantiate three general paradigms—*Attribution Reasoning* (McArthur (1972); Hastie (1984)), *Counterfactual Reasoning* (Pearl (2022)), and *Causal Mediation Reasoning* (Imai et al. (2010); Pearl (2014))—as guiding principles for CoT generation. For specialized medical domains, we incorporate *Evidence-Based Medicine* (Sackett et al. (1996); Guyatt et al. (1992)), *Precision Medicine* (Council et al. (2011); Collins & Varmus (2015)), and *Integrative Medicine* (Maizes et al. (2009); Snyderman & Weil (2002)) paradigms to guide complex diagnostic and therapeutic decision-making.

ParadigmCoT performs paradigm-aware policy optimization within a reinforcement learning framework, enabled by a novel Automated Paradigm Localization and Evaluation module. The module first identifies paradigm-initiation sentences using targeted keywords, and then evaluates the adherence of the subsequent q reasoning steps to the corresponding paradigm in light of the preceding context. The resulting paradigm-alignment scores, combined with step-wise factuality and answer correctness, provide fine-grained supervision through an Advantage Adjustment Mechanism. Together, these signals form a dense reward structure that steers policy optimization toward more reliable, paradigm-aligned reasoning. The overall framework is illustrated in Fig. 2. Across three LLM backbones and diverse general and medical benchmarks, ParadigmCoT achieves substantial gains over base models (up to 40.2% on general tasks and 34.5% on medical tasks), and outperforms RL-based baselines by 2.5–3.4% and 5.8–7.7% on average, These results highlight paradigm-level guidance as a promising direction for improving the quality and reliability of LLM-generated CoT, thereby advancing more reliable AI. Our main contributions are as follows:

- We propose paradigm-guided CoT reasoning for LLMs, motivated by the limitation that conventional CoT makes reasoning explicit but offers limited principle-level guidance over the full inference trajectory. By anchoring intermediate steps to general and domain-specific instructive paradigms, it steers the overall reasoning trajectory toward more principled and reliable inference.
- We propose ParadigmCoT, a general and flexible framework that injects principle-level guidance into CoT via paradigm-aware policy optimization. ParadigmCoT uses an automated paradigm localization and evaluation module, together with a fine-grained advantage adjustment mechanism, to jointly optimize paradigm adherence, factuality, and answer accuracy for reliable reasoning.
- Extensive experiments on general and medical benchmarks with three LLM backbones show that ParadigmCoT substantially improves over base models, with gains of up to 40.2% and 34.5%, respectively, and outperforms RL-based baselines by 2.5%–3.4% and 5.8%–7.7% on average, demonstrating robust effectiveness in both general and medical domains.

2 Related Work

CoT and Structured Prompting Methods—Lacking Paradigm-Level Guidance. CoT prompting (Wei et al. (2022); Kojima et al. (2022)) and its extensions, such as self-consistency (Wang et al. (2022)) and sub-question decomposition (Zhou et al. (2022)), have markedly enhanced LLMs' ability to tackle complex reasoning tasks by enabling stepwise inference. However, these approaches often prioritize linguistic plausibility over logical rigor (Turpin et al. (2023); Lanham et al. (2023)), occasionally generating plausible-sounding but logically disjointed justifications. To further enforce systematicity, methods such as Program-of-Thought (Chen et al. (2023)), Tree of Thoughts (Yao et al. (2023)), and Graph of Thoughts (Besta et al. (2024)) introduce symbolic or structural scaffolds. These innovations primarily expand the diversity of reasoning paths in a structured manner, enhancing performance on creative and strategic tasks; however, they lack the principled, paradigm guidance required to steer the entire CoT process toward systematic and logically consistent reasoning.

Process Supervision Methods—Lacking Holistic Paradigm Guidance. In parallel, process-based supervision frameworks such as Math-Shepherd (Wang et al. (2024)) and FSPO (Li & Ng) provide more fine-grained signals by evaluating statements or step-wise outputs during generation. These works have made significant contributions to densifying feedback and mitigating reward exploitation (Uesato et al. (2022); Huang et al. (2024)), representing an important advance beyond traditional outcome-based objectives. However, most of these frameworks focus primarily on isolated, local verification of factual correctness or answer-level accuracy (Tarek & Beheshti (2025); Lee et al. (2024)), yet lack principle-level, paradigm-based supervision that guides the entire CoT process. This highlights the urgent need for a general framework capable of providing such holistic paradigm guidance throughout reasoning chains, thereby enabling more reliable LLM reasoning.

3 Preliminary

We review the reinforcement learning framework for language modeling and the Group Relative Policy Optimization (GRPO) algorithm, which serves as the optimization foundation for ParadigmCoT.

3.1 Reinforcement Learning Formulation

The process of stepwise language generation can be cast as a Markov Decision Process at the token level, denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \pi_\theta)$. At timestep t, the state $s_t \in \mathcal{S}$ is constructed by concatenating the input prompt x with the previously generated tokens $y_{<t} = [o_1, \ldots, o_{t-1}]$. The policy $\pi_\theta(\cdot \mid s_t)$, parameterized by θ, chooses the next token $o_t \in \mathcal{A}$ from the vocabulary, resulting in a deterministic transition to $s_{t+1} = [s_t, o_t]$. Let $\mathcal{Y}^*$ denote the set of correct output sequences for the given input x. Most methods employ a binary reward function $\mathcal{R}$ for each output y (Ladosz et al. (2022); Guo et al. (2025)). In this paper, We follow a prevalent format: $\mathcal{R}(y) = 1$ if $y \in \mathcal{Y}^*$, and $\mathcal{R}(y) = 0$ otherwise. The expected reward achieved by the policy π_θ is given by:

$$\mathcal{J}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[\mathcal{R}(y)] = \sum\nolimits_y \mathcal{R}(y)\, \pi_\theta(y|x). \tag{1}$$

With a binary reward, the expected return reflects the likelihood that the policy π_θ generates a correct sequence. To maximize this probability, standard reinforcement learning applies policy gradient algorithms, with the objective gradient formulated as:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[\mathcal{R}(y) \nabla_\theta \log \pi_\theta(y|x) \right]. \tag{2}$$

3.2 Group Relative Policy Optimization

Our method builds on Group Relative Policy Optimization (GRPO) (Shao et al. (2024)), an efficient extension of Proximal Policy Optimization (PPO) (Schulman et al. (2017)) that does not require a value function. Specifically, for each input x, GRPO samples a group of responses $\{y_1, ..., y_G\}$, and computes their rewards $\{\mathcal{R}_1, ..., \mathcal{R}_G\}$. The advantage for every token in sequence y_i is assigned as $A_{i,t} = (R_i - \text{mean}(R_1, \ldots, R_G))/\text{std}(R_1, \ldots, R_G)$. The objective for GRPO can be expressed as:

$$\begin{aligned} &\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{y \sim \pi(\cdot|x)} \frac{1}{G} \\ &\sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left\{ \min \left[\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})} A_{i,t},\ \text{clip} \left(\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})}, 1-\epsilon, 1+\epsilon \right) A_{i,t} \right] \right\}, \end{aligned} \tag{3}$$

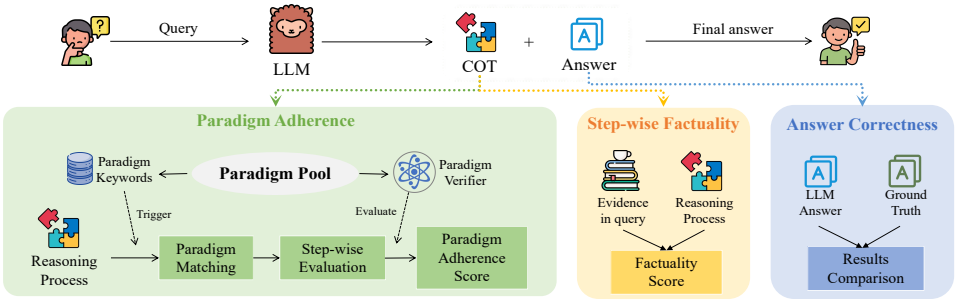


Figure 2: Overview of ParadigmCoT. CoT reasoning is evaluated using step-wise paradigm adherence, factuality, and final answer correctness, and these signals jointly guide policy optimization toward more reliable and paradigm-aligned reasoning.

4 METHOD

We propose a paradigm-aware policy optimization framework, as illustrated in Fig.2. Section 4.1 details paradigm guidance for general and medical domains. Section 4.2 presents the composite reward function integrating step-wise paradigm adherence, factuality, and answer correctness. Section 4.3 describes the paradigm-aware advantage adjustment for paradigm-aligned reasoning

4.1 PARADIGM GUIDANCE

ParadigmCoT bridges the gap between unconstrained LLM-generated CoT and principled reasoning by injecting paradigms as process-level guidance. As summarized in Table 1, both general paradigms (e.g., attribution, counterfactual, causal mediation reasoning) and clinical paradigms in medicine (e.g., evidence-based, precision and integrative Medicine) are clearly delineated and further detailed through explicit instruction templates (see Appendix A). These templates provide clear and easy-to-follow guidance for the LLM, ensuring that the model's intermediate outputs are anchored within instructive and principled paradigms.

Table 1: Summary of paradigms instantiated in ParadigmCoT for general and medical domains, including their explanations and applications, as well as key theoretical references.

Paradigm	**Paradigm Explanation and Application**	**Theoretical Ref.**
General Paradigms		
Attribution Reasoning	Evaluate whether outputs are supported by the provided evidence or information sources, ensuring that inference is grounded in explicit facts.	McArthur (McArthur (1972)); Hastie (Hastie (1984))
Counterfactual Reasoning	Analyze how changing certain input conditions may affect the outcome and reveal which factors most strongly influence results.	Pearl (Pearl (2022))
Causal Mediation	Decompose causal relationships to identify and explain intermediate variables or processes that mediate the effect from cause to outcome.	Imai et al. (Imai et al. (2010)); Pearl (Pearl (2014))
Clinical Paradigms in Medicine		
Evidence-Based Medicine	Use high-quality clinical evidence and guideline strength to determine medical decisions and recommendations.	Sackett et al. (Sackett et al. (1996)); Guyatt et al. (Guyatt et al. (1992))
Precision Medicine	Personalize diagnosis and treatment based on individual patient features, including genomics, age, comorbidities, and other relevant factors.	NRC (Council et al. (2011)); Collins (Collins & Varmus (2015))
Integrative Medicine	Combine biomedical, psychological, and lifestyle factors to make holistic clinical assessments, especially for chronic or elderly cases.	Maizes et al. (Maizes et al. (2009)); Snyderman (Snyderman & Weil (2002))

4.2 PARADIGM ADHERENCE REWARD FUNCTION DESIGN

The core of reinforcement learning is to design reward functions that promote logical reasoning and suppress errors, guiding the policy toward improved performance. In ParadigmCoT, we introduce three complementary reward signals: step-wise paradigm adherence and input-grounded factuality as process-level supervision, and final answer correctness as outcome-level supervision.

4.2.1 AUTOMATED PARADIGM LOCALIZATION AND EVALUATION

ParadigmCoT employs an *automated paradigm localization and evaluation* module to supervise paradigm adherence in the reasoning chain. Given a reasoning chain $y = [z_1, \ldots, z_N]$ of N sentences, where each z_j denotes the j-th sentence, the module scans the chain to identify candidate paradigm-triggering sentences by matching paradigm-specific keywords (e.g., "because", "due to" for attribution reasoning; see Appendix B for the full list).

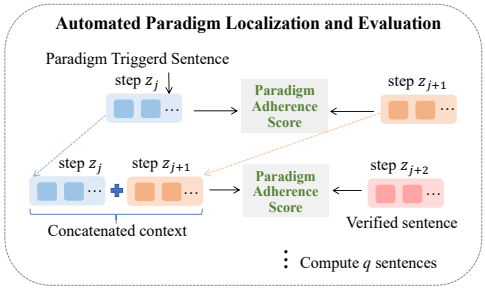


Figure 3: The module computes paradigm-adherence scores for the subsequent q reasoning steps based on the preceding context.

Once a paradigm-triggering sentence z_j is identified, we evaluate its following q steps ($k = 1, \ldots, \min(q, N-j)$). For each step, we concatenate the trigger and preceding reasoning sentences into $C_{j,k} = \text{Concat}(z_j, z_{j+1}, \ldots, z_{j+k-1})$, and score the adherence of the next sentence z_{j+k} to this context (see Fig. 3). If the score exceeds a threshold τ, both z_{j+k} and z_j are rewarded. This validates candidate triggers through subsequent semantic adherence and ensures that paradigm rewards reflect coherent extensions of the CoT trajectory.

$$\mathcal{R}_{\text{para}}(z_{j+k}) = \begin{cases} 1, & \text{if } S(C_{j,k}, z_{j+k}) > \tau, \\ 0, & \text{otherwise,} \end{cases} \qquad \mathcal{R}_{\text{para}}(z_j) = \max_{k=1}^{\min(q,N-j)} \mathcal{R}_{\text{para}}(z_{j+k}).$$

where $S(\cdot,\cdot)$ is computed by HHEM-2.1 (Hughes et al. (2023)), an external verifier that reduces self-reinforcing bias relative to using the same generative LLM as a judge and is efficient enough for training-time reward computation. We use $\tau = 0.35$ (see subsection 5.4) and set $q = 4$ empirically. To keep the paradigm reward bounded and prevent inflation from keyword stacking, we assign at most one paradigm reward per sentence: triggers in different sentences are evaluated independently with separate q-step windows, while multiple triggers within the same sentence are resolved by a first-match rule. The reliability of this module is analyzed in subsection 5.4.

4.2.2 FACTUALITY AND ANSWER REWARDS

Factuality Rewards. Adhering to paradigms provides strategic guidance for the entire CoT reasoning process, while maintaining faithfulness to the input context keeps reasoning grounded. Therefore, we introduce a step-level factuality reward inspired by FSPO (Li & Ng). Specifically, each sentence z_j in the reasoning chain is compared against the evidence set $\mathcal{K}$ (i.e., the full input context) by computing an entailment score $S_{\text{fact}}(z_j, \mathcal{K})$ using HHEM-2.1 (Hughes et al. (2023)), with a threshold of 0.5 (following FSPO). The reward is then set as:

$$\mathcal{R}_{\text{factual}}(z_j) = \begin{cases} 1, & \text{if } S_{\text{fact}}(z_j, \mathcal{K}) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Answer Rewards. A binary reward on the final answer is widely employed for its effectiveness and computational simplicity (Guo et al. (2025); Xie et al. (2025)). Given an input x, the model generates an output y that contains both the reasoning process and a final answer. A rule-based reward function is employed to evaluate an answer's correctness:

$$\mathcal{R}_{\text{answer}}(y) = \begin{cases} 1, & \text{if } y \text{ fully matches the ground truth} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The final reward for the reasoning process aggregates answer correctness and process-level signals:

$$R_{\text{final}}(y) = R_{\text{answer}}(y) + \frac{1}{N} \sum_{j=1}^{N} [R_{\text{factual}}(z_j) + R_{\text{para}}(z_j)] \tag{6}$$

This composite reward scheme provides dense feedback, encouraging not only answer correctness, but also reliable reasoning that is both guided by paradigms and faithful to input evidence.

4.3 PARADIGM-AWARE POLICY OPTIMIZATION

While Group Relative Policy Optimization (GRPO) (Shao et al. (2024)) effectively balances reward among sampled responses, it does not incorporate process-level paradigm alignment within intermediate reasoning steps. To address this, we extend GRPO by introducing a paradigm-aware token-level advantage adjustment mechanism: we re-weight each token's advantage according to its alignment

Table 2: Advantage adjustment matrix. Left: correct answer; right: incorrect answer.

$\mathcal{R}_{\text{answer}}$	$\mathcal{R}_{\text{factual}}$	$\mathcal{R}_{\text{para}}$	$A^*_{i,t}$	Case	$\mathcal{R}_{\text{answer}}$	$\mathcal{R}_{\text{factual}}$	$\mathcal{R}_{\text{para}}$	$A^*_{i,t}$	Case
1	1	1	$1.2\bar{A}_i$	Synergistic success	0	0	0	$-1.2\bar{A}_i$	Primitive error
1	1	0	$1.0\bar{A}_i$	Factual-only success	0	0	1	$-1.0\bar{A}_i$	Paradigm misuse
1	0	0	$0.8\bar{A}_i$	Spurious correct	0	1	0	$-0.8\bar{A}_i$	Factual-only failure
1	0	1	$-0.8\bar{A}_i$	Paradigm-only success	0	1	1	$0.8\bar{A}_i$	Paradigm persistence

with outcome- and process-level signals. Specifically, for each token $o_{i,t}$ in reasoning step z_j, we compute $A^*_{i,t}$ by first taking the absolute value of the GRPO advantage, $\bar{A}_i = |A_i|$, and then applying the reward–penalty multiplier in Table 2 according to $(R_{\text{answer}}(y), R_{\text{factual}}(z_j), R_{\text{para}}(z_j))$.

Table 2 is designed as follows: the strongest reward/penalty multipliers (± 1.2) are used when all criteria are fully met or missed. For correct answers, positive multipliers are assigned to both factual-only and spurious cases ($1.0, 0.8$), as not all examples require paradigm-guided reasoning. For incorrect answers, paradigm-aligned but wrong reasoning ("Paradigm misuse") and factual-only failure are penalized with negative multipliers ($-1.0, -0.8$). "Paradigm-only success" is also penalized when it is not faithful to the input, discouraging unwarranted paradigm use. If both factual and paradigm alignment are present despite an incorrect answer ("Paradigm persistence"), a mild positive multiplier is given to preserve high-quality CoT. The rationale for the penalty–reward magnitude (± 0.2 around 1) is further analyzed in subsection 5.4.

This adjustment provides fine-grained gradient signals based on both outcome- and process-level supervision, strengthening updates for paradigm-aligned and faithful reasoning steps while suppressing spurious or misaligned ones, which guides the LLM toward principled and logically consistent reasoning. Based on the adjusted advantage, our policy objective is formulated as follows:

$$\mathcal{J}_{\text{ParadigmCoT}}(\theta) = \mathbb{E}_{y\sim\pi(\cdot|x)} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left\{ \min \left[\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})} A^*_{i,t},\ \text{clip}\left(\frac{\pi_\theta(o_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|x, y_{i,<t})}, 1-\epsilon, 1+\epsilon \right) A^*_{i,t} \right] \right\}, \tag{7}$$

5 Experiment

5.1 Experimental Setup

Datasets and Metrics. We train and evaluate ParadigmCoT separately in the general and medical domains with their corresponding paradigm sets. For the **general domain**, training uses subsets of HotpotQA (Yang et al. (2018)), 2K examples from 2WikiMultiHopQA (Ho et al. (2020)), and 8K problems from SimpleRL (Zeng et al. (2025)). Evaluation includes mathematical reasoning benchmarks, i.e., GSM8K (Cobbe et al. (2021)), MATH-500 (Hendrycks et al. (2021c)), AIME 2024 (Codeforces), and AIME 2025 (Zhang & Math-AI (2025)), as well as hallucination benchmarks, i.e., TruthfulQA (Lin et al. (2021)), HaluEval-QA (Li et al. (2023)), and HalluQA (Cheng et al. (2023)). For the **medical domain**, we train on MedAesQA (Gupta et al. (2025; 2024)) and PubMedQA (Jin et al. (2019)), and evaluate on MMLU-Med (Hendrycks et al. (2021b;a)), MedQA (Jin et al. (2021)), and two carefully annotated, high-quality in-house clinical benchmarks: **DiagTreat** for diagnosis-treatment consistency and **MedJudge** for medication appropriateness, detailed in Appendix C. We report Pass@1 accuracy on mathematical reasoning benchmarks, accuracy on medical benchmarks and HaluEval-QA, and the fraction of judge-labeled truthful outputs on TruthfulQA and HalluQA.

Baselines. We compare against five groups of competitive baselines: (1) **API-based models**: DeepSeek-V3 (Liu et al. (2024)), DeepSeek-R1 (Guo et al. (2025)), OpenAI GPT-o1, and GPT-5.2; (2) **Reasoning models**: QwQ-32B (Team (2025)), Distill-Qwen-7B/14B/32B, and Distill-Llama-8B (Guo et al. (2025)), fine-tuned on 800K long CoT samples synthesized from DeepSeek-R1; (3) **Base models**: Qwen2.5-7B-Base/Instruct and Llama3.1-8B-Instruct; (4) **Structured prompting models**: standard CoT (Wei et al. (2022)), Tree-of-Thought (ToT) (Yao et al. (2023)), and Graph-of-Thought (GoT) (Besta et al. (2024)), all instantiated with Qwen2.5-7B-Instruct for a consistent prompting-based comparison; and (5) **RL-based models**: FSPO (Li & Ng), DAPO (Yu et al. (2026)), and GSPO (Zheng et al. (2025)), evaluated across the three aforementioned base models.

Implementation Details. All training-based methods, including ParadigmCoT and the RL baselines, are trained under the same data and configuration; RL training is conducted with verl (Sheng et al.

(2025)) on 4×H100 80G GPUs for about 24 hours per run, using 8 prompts and 8 rollouts per prompt per batch. We use a mini-batch size of 1,024, one epoch, learning rate 4e-7, rollout temperature 1.0, KL coefficient 1e-3, clipping ratio 0.2, and a 2,048-token maximum length. DAPO and GSPO use official implementations; FSPO results are taken from the original report when available and otherwise reproduced under the same setup. For evaluation, ParadigmCoT and directly compared baselines use greedy decoding with a 2,048-token maximum response length, while API-based and reasoning models use default generation settings. Prompts are provided in Appendix D.

5.2 Main Results in the General Domain

Table 3 shows that ParadigmCoT achieves consistent improvements across mathematical and hallucination-related reasoning tasks after integrating three general instructive paradigms into the CoT process. It substantially outperforms base and structured prompting models, and further improves over representative RL-based baselines by 2.50%–3.12% on mathematical reasoning and 2.50%–3.44% on hallucination detection. These results support the contribution of ParadigmCoT: aligning intermediate CoT steps with instructive paradigms provides principle-level guidance over the reasoning trajectory and promotes more systematic, logically consistent reasoning. They further show that such paradigms can be integrated via policy optimization, yielding consistent gains on both mathematical and hallucination-related benchmarks. Finally, ParadigmCoT consistently improves performance across three different backbones, indicating robust effectiveness under varying LLMs settings.

5.3 Main Results in Medicine

Table 4 reports the performance of ParadigmCoT on medical QA and clinical reasoning benchmarks across three backbone models. ParadigmCoT clearly outperforms base, structured prompting, and RL-based methods. Averaged across all backbones, ParadigmCoT improves over the three RL-based baselines by 5.82%, 7.71%, 4.31%, and 6.16% on the four datasets, respectively, while matching much larger distilled reasoning models with only 7B/8B backbones.

These gains validate the effectiveness of paradigm-guided supervision in medical reasoning, showing that ParadigmCoT promotes domain-specific and systematic reasoning trajectories. For medical QA, it helps align relevant medical knowledge with the question context; for clinical reasoning, the injected paradigms guide the model to connect symptoms, diagnoses, treatments, and decision criteria within the CoT process, leading to more comprehensive clinical judgments. A case study illustrating the paradigm-guided process is provided in Appendix E. Consistent improvements across different base models indicate robust performance, highlighting ParadigmCoT's potential for more reliable domain-specific deployment of LLMs.

Table 4: Evaluation on medical QA and clinical reasoning tasks. **Bold** marks the best result among base, structured prompting, and RL-based models for each backbone.

Model	MED-QA	MMLU-MED	Diag	Treat	Med Judge
API-based Models					
DeepSeek-V3	85.5	88.0	73.7	86.4	
DeepSeek-R1	89.1	91.7	75.6	88.3	
GPT-o1-1217	88.7	91.3	74.8	86.3	
GPT-5.2	90.5	92.1	76.7	89.5	
Reasoning Models					
QwQ-32B	85.2	89.2	71.4	83.2	
R1-Distill-Qwen-32B	87.6	88.9	72.9	85.7	
R1-Distill-Qwen-14B	85.0	87.7	71.1	84.2	
R1-Distill-Qwen-7B	84.3	86.2	70.3	82.9	
R1-Distill-Llama-8B	82.2	85.1	68.4	80.4	
Base Models					
Qwen2.5-7B-Base	53.1	69.8	59.8	76.2	
Qwen2.5-7B-Instruct	57.4	74.3	62.5	79.4	
Llama3.1-8B-Instruct	50.7	69.8	60.5	77.4	
Structured Prompting					
CoT (Qwen-Instruct)	59.8	77.8	66.2	79.9	
ToT (Qwen-Instruct)	61.2	77.4	69.0	78.9	
GoT (Qwen-Instruct)	62.3	78.1	67.5	78.6	
RL-based Models					
GSPO (Qwen-Base)	79.4	78.9	66.0	79.5	
FSPO (Qwen-Base)	81.4	82.6	69.4	81.2	
DAPO (Qwen-Base)	80.3	81.1	69.3	81.3	
ParadigmCoT (Qwen-Base)	**85.4**	**87.2**	**70.5**	**85.4**	
GSPO (Qwen-Instruct)	82.0	82.2	68.8	79.5	
FSPO (Qwen-Instruct)	80.1	83.2	70.7	82.9	
DAPO (Qwen-Instruct)	81.1	82.5	71.2	83.3	
ParadigmCoT (Qwen-Instruct)	**86.6**	**89.4**	**72.6**	**87.3**	
GSPO (Llama-Instruct)	76.7	76.5	62.0	75.3	
FSPO (Llama-Instruct)	78.7	77.3	64.4	76.5	
DAPO (Llama-Instruct)	79.5	78.6	68.3	77.9	
ParadigmCoT (Llama-Instruct)	**85.2**	**87.5**	**73.2**	**84.9**	

5.4 Further Analysis

Ablation Analysis. We compare ParadigmCoT with four variants: **w/o Factuality**, which removes the factuality reward; **w/o Paradigm**, which removes both the paradigm prompt and paradigm reward; **w/o Paradigm Reward (+paradigm prompt)**, which keeps the paradigm prompt but removes the

Table 3: Performance comparison across math reasoning and hallucination benchmarks. **Bold** indicates the best results among base, structured prompting, and RL-based models for each backbone.

Model	Math Reasoning Tasks				Hallucination Tasks		
	GSM8K	MATH -500	AIME24 (Pass@1)	AIME25 (Pass@1)	TruthfulQA	HaluEval-QA	HalluQA
API-based Models							
DeepSeek-V3	89.3	90.2	39.2	26.6	53.0	70.6	59.1
DeepSeek-R1	92.0	97.3	79.8	66.7	45.2	51.8	51.1
GPT-o1-1217	92.3	96.4	79.2	76.7	62.4	70.2	56.0
GPT-5.2	98.1	97.7	100	100	78.2	75.5	59.3
Reasoning Models							
QwQ-32B	88.6	89.8	79.5	56.7	23.2	29.6	23.7
R1-Distill-Qwen-32B	87.4	94.3	72.6	53.3	19.7	33.5	26.9
R1-Distill-Qwen-14B	85.1	93.9	69.7	50.0	18.2	24.8	20.2
R1-Distill-Qwen-7B	84.3	92.8	55.5	33.3	6.9	11.6	3.1
R1-Distill-Llama-8B	82.1	89.1	50.4	26.7	8.8	14.6	4.6
Base Models							
Qwen2.5-7B-Base	65.2	35.7	3.3	3.3	38.2	48.0	39.5
Qwen2.5-7B-Instruct	73.2	51.6	6.7	3.3	36.7	38.4	32.0
Llama3.1-8B-Instruct	77.5	33.1	6.7	0.0	26.4	36.7	12.2
Structured Prompting Models							
CoT (Qwen-Instruct)	77.3	52.0	6.7	3.3	37.5	38.8	33.3
ToT (Qwen-Instruct)	79.5	57.6	10.0	9.0	39.4	42.8	41.6
GoT (Qwen-Instruct)	74.2	54.8	10.0	3.3	37.7	41.2	33.1
RL-based Models							
GSPO (Qwen-Base)	88.5	71.6	14.4	11.3	55.2	68.9	50.5
FSPO (Qwen-Base)	89.5	75.5	16.7	13.3	56.4	71.3	52.0
DAPO (Qwen-Base)	90.1	74.4	**20.0**	12.6	57.8	71.1	53.1
ParadigmCoT (Qwen-Base)	**91.3**	**75.9**	16.7	**14.6**	**60.2**	**72.6**	**54.3**
GSPO (Qwen-Instruct)	86.7	68.2	15.4	6.7	55.4	62.2	48.4
FSPO (Qwen-Instruct)	89.4	74.7	20.0	13.3	52.0	61.7	47.2
DAPO (Qwen-Instruct)	91.1	73.6	20.0	13.0	52.2	61.7	49.8
ParadigmCoT (Qwen-Instruct)	**92.2**	**78.1**	**23.3**	**14.0**	**56.4**	**66.8**	**52.2**
GSPO (Llama-Instruct)	86.4	67.4	13.2	4.8	39.3	66.6	40.5
FSPO (Llama-Instruct)	86.2	68.3	13.3	6.7	41.1	67.1	42.0
DAPO (Llama-Instruct)	87.4	69.6	12.2	6.7	41.7	66.2	41.6
ParadigmCoT (Llama-Instruct)	**89.1**	**69.8**	**16.7**	**9.0**	**44.1**	**68.2**	**42.7**

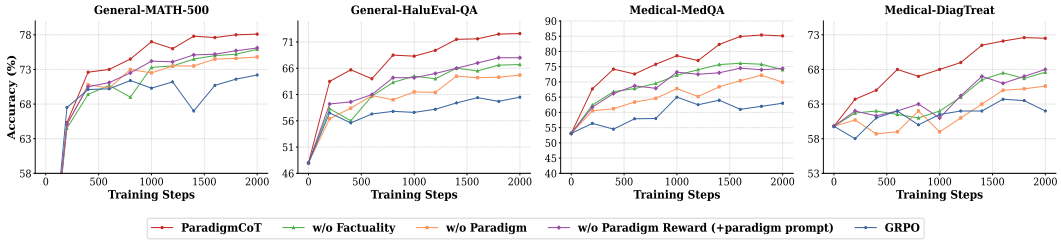


Figure 4: Ablation analysis of Pass@1 accuracy over training steps on general and medical benchmarks.

paradigm reward; and standard **GRPO**. As shown in Fig. 4, **w/o Paradigm** shows a substantial performance drop, especially on tasks requiring complex reasoning, confirming the importance of paradigm guidance in steering the CoT process. **w/o Paradigm Reward (+paradigm prompt)** yields limited gains, indicating that most improvements come from paradigm-aware policy optimization rather than paradigm prompts. Removing the factuality reward causes moderate degradation, while GRPO performs worst overall. Together, these ablations validate the ParadigmCoT design and highlight the key role of paradigm-aware process supervision in improving performance.

Analysis of Paradigm Localization and Evaluation. We first compare paradigm-trigger and assigned reward densities. As shown in Table 5, triggers occur more often than rewards, indicating that semantic-adherence checks filter many keyword hits and retain rewards that reflect coherent extensions of the reasoning trajectory. We then manually annotate 50 rewarded CoTs per dataset and observe high precision (86.91%–91.77%), suggesting that most rewarded trajectories exhibit valid paradigm-guided reasoning. We also annotate the missed-paradigm rate among incorrect responses

without paradigm rewards, i.e., cases where an appropriate paradigm appears applicable but is not detected by the localization module. The result suggest limited implicit paradigm guidance, supporting that the module provides reliable paradigm supervision while remaining efficient for online reward computation.

Table 5: Paradigm trigger and actual reward density, with manually annotated reward precision and missed-paradigm rate.

Datasets	Trigger Density	Paradigm Reward Density	Reward Precision	Missed Paradigm Rate
GSM8K	45.60%	31.40%	91.77%	5.40%
MATH-500	47.40%	32.50%	87.35%	11.51%
MedQA	48.70%	28.30%	86.91%	8.71%
DiagTreat	47.60%	29.80%	88.33%	11.68%

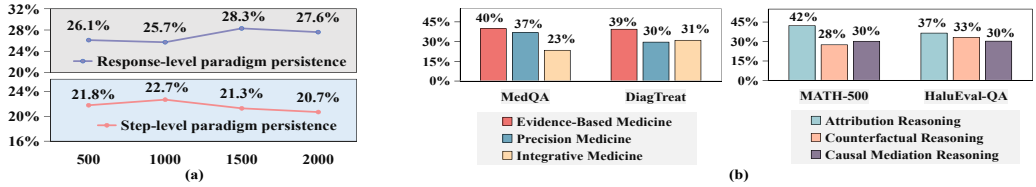


Figure 5: (a) Proportions of response- and step-level paradigm persistence in incorrect responses across training steps. (b) Paradigm distributions of correct predictions across four datasets.

Paradigm Activation Dynamics. To assess whether paradigm rewards induce process-level reward hacking, we track paradigm occurrence rates in incorrect responses across training steps at both the response and step levels: the former is the fraction of incorrect responses containing at least one paradigm-guided reasoning step, and the latter is the fraction of paradigm-adherent steps among all reasoning steps. As shown in Fig. 5(a), these paradigm persistence rates remain broadly stable over training (with a slight decrease at the step level), suggesting no clear reward hacking via the accumulation of paradigm-adherent steps.

Paradigm Distributions. We examine the paradigms activated in correctly predicted examples, as shown in Fig. 5 (b). ParadigmCoT activates all paradigms across general and medical tasks, suggesting their complementary roles, while their proportions vary by dataset and reflect task-specific reasoning demands. For example, MedQA, as a knowledge-intensive QA benchmark, favors evidence-based and precision-medicine paradigms, whereas DiagTreat relies more on evidence-based and integrative-medicine paradigms, reflecting its emphasis on comprehensive consideration.

Hyperparameter Analysis. Fig. 6 studies two key hyperparameters. The **penalty-reward granularity** adjusts advantage scaling relative to the default GRPO coefficient of 1, enabling finer control of the learning signal (Table 2). Performance on three datasets improves from 0.0 to 0.2 and then declines, showing that moderate fine-grained adjustment is beneficial while excessive scaling can disturb optimization. The **paradigm adherence threshold** controls paradigm reward activation, with all three benchmarks peaking at 0.35. This threshold balances meaningful semantic adherence and sufficient paradigm activation; larger values become overly restrictive. We therefore set the granularity to 0.2 and the threshold to 0.35 in all experiments.

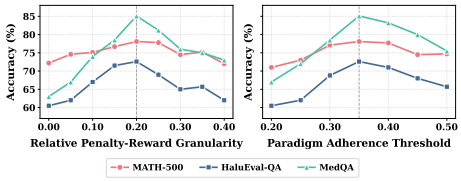


Figure 6: Hyperparameter analysis of penalty–reward granularity and paradigm-adherence threshold.

6 Conclusion and Limitations

We propose ParadigmCoT for paradigm-guided CoT reasoning in LLMs, motivated by the limited principle-level guidance of conventional CoT over full inference trajectories. By anchoring intermediate steps to instructive paradigms, ParadigmCoT steers CoT trajectories toward more systematic and logically consistent generation. Across general and medical benchmarks and three LLM backbones, ParadigmCoT clearly improves over base models and competitive baselines, highlighting paradigm-level guidance as a promising direction for more reliable LLM reasoning.

ParadigmCoT has two limitations. First, although many domains have well-established instructive paradigms, applying it to domains with less explicit reasoning paradigms may require paradigm induction and adaptation, including the construction of suitable templates and keywords. Second, our localization module uses keyword-triggered detection for practical and controllable CoT process supervision, which may miss some implicit paradigm usage. Future work will explore learned paradigm localization and evaluation methods to improve paradigm recall and adaptability.

AI USE STATEMENT

Large language models (LLMs) were used solely to assist with the writing and polishing of this manuscript. Specifically, we used LLMs for grammar checking, sentence rephrasing, improving readability, and enhancing the clarity and flow of the text.

ETHICS STATEMENT

This work proposes ParadigmCoT to improve the reasoning reliability of LLMs by reducing reasoning-path drift and logical inconsistency through paradigm-guided CoT process supervision. Our experiments cover standard general-domain benchmarks, public medical QA benchmarks, and two in-house clinical reasoning benchmarks for evaluating diagnosis–treatment consistency and medication appropriateness. Regarding data ethics and privacy, the public datasets used in this work are employed only for research evaluation and are properly cited. The two in-house clinical benchmarks, DiagTreat and MedJudge, consist of fully synthetic clinical vignettes and contain no patient-derived data. No hospital electronic health records, patient records, or identifiable personal information were accessed or used. All cases were generated de novo and reviewed by clinicians for clinical plausibility and label consistency. Therefore, this work does not involve human-subjects data, and IRB approval or waiver is not applicable.

Because medical reasoning is a high-stakes application area, incorrect model outputs may pose risks if misused, including unsafe diagnostic suggestions, missed contraindications or drug–drug interactions, inappropriate medication recommendations, or plausible but unreliable rationales. Therefore, our medical results should be understood only as controlled benchmark evaluations of reasoning behavior, not as evidence that ParadigmCoT is ready for clinical deployment. We do not advocate using ParadigmCoT for autonomous clinical decision-making. Any real-world medical use would require additional validation in clinical workflows, oversight by qualified medical professionals, and safety mechanisms to prevent unsupported or harmful recommendations.

REPRODUCIBILITY STATEMENT

We provide the main implementation details in Section 5.1, including the datasets, backbone models, training setup, decoding settings, hyperparameters, compute resources, and baseline configurations used in our experiments. Details of the paradigm templates and paradigm-triggering keywords are provided in Appendices A and B, and descriptions of the in-house clinical benchmarks are provided in Appendix C. Prompt templates for all evaluated tasks are included in Appendix D. These materials are intended to facilitate reproduction of the training and evaluation pipeline. Additionally, we will provide an anonymized link to the code repository.

REFERENCES

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023.

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

MAA Codeforces. American invitational mathematics examination-aime 2024, 2024.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

National Research Council, Division on Earth, Life Studies, Board on Life Sciences, and Committee on A Framework for Developing a New Taxonomy of Disease. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. 2011.

Sebastian Farquhar, Vikrant Varma, David Lindner, David Elson, Caleb Biddulph, Ian Goodfellow, and Rohin Shah. Mona: Myopic optimization with non-myopic approval can mitigate multi-step reward hacking. In *Forty-second International Conference on Machine Learning*.

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.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. Overview of trec 2024 biomedical generative retrieval (biogen) track. *TREC*, 2024.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1): 1035, 2025.

Gordon Guyatt, John Cairns, David Churchill, Deborah Cook, Brian Haynes, Jack Hirsh, Jan Irvine, Mark Levine, Mitchell Levine, Jim Nishikawa, et al. Evidence-based medicine: a new approach to teaching the practice of medicine. *jama*, 268(17):2420–2425, 1992.

Reid Hastie. Causes and effects of causal attribution. *Journal of personality and social psychology*, 46(1):44, 1984.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021c.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. Key-point-driven data synthesis with its enhancement on mathematical reasoning. *arXiv preprint arXiv:2403.02333*, 2024.

Simon Hughes, Minseok Bae, and Miaoran Li. Vectara Hallucination Leaderboard, November 2023. URL https://github.com/vectara/hallucination-leaderboard.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*, 2024.

Junyi Li and Hwee Tou Ng. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. *URL https://arxiv. org/abs/2109.07958*, 1, 2021.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2908–2921, 2023.

Victoria Maizes, David Rakel, and Catherine Niemiec. Integrative medicine and patient-centered care. *Explore*, 5(5):277–289, 2009.

Leslie A McArthur. The how and what of why: Some determinants and consequences of causal attribution. *Journal of personality and Social Psychology*, 22(2):171, 1972.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and causal inference: the works of Judea Pearl*, pp. 317–372. 2022.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't, 1996.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Ralph Snyderman and Andrew T Weil. Integrative medicine: bringing medicine back to its roots. *Archives of Internal medicine*, 162(4):395–397, 2002.

Mirza Farhan Bin Tarek and Rahmatollah Beheshti. Reward hacking mitigation using verifiable composite rewards. *arXiv preprint arXiv:2509.15557*, 2025.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 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.

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

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient. `https://hkust-nlp.notion.site/simplerl-reason`, 2025. Notion Blog.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime). *Mathematical Association of America (MAA)*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL `https://arxiv.org/abs/2507.18071`.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

APPENDIX

A PARADIGM GUIDANCE TEMPLATES

To support principled CoT reasoning, ParadigmCoT systematically integrates both general and medical domain paradigms into model prompts. At the end of each instruction, the following paradigms are appended and made explicit to the LLM as stepwise process scaffolds:

General Reasoning Paradigms If applicable, utilize one of the following reasoning paradigms to enhance the accuracy of your solution:

- **Attributional Reasoning:** Evaluate whether each generated claim can be strictly substantiated by underlying factual evidence.
- **Causal Mediated Reasoning:** Identify and intervene through intermediary variables between the input factors and the final result to refine the internal logic flow.
- **Counterfactual Reasoning:** Analyze the causal drivers of a prediction by exploring "what if" scenarios, highlighting how specific input variations would alter the outcome.

Clinical Reasoning Paradigms in Medicine If applicable, utilize one of the following reasoning paradigms to enhance the accuracy of your solution:

- **Evidence-Based Medicine:** Base all diagnostic and treatment suggestions on the latest and most reliable clinical evidence.
- **Precision Medicine:** Tailor recommendations according to the patient's unique features (such as age, genetic background, comorbidities, or lab findings) to maximize individual benefit.
- **Integrative Medicine:** Integrate biomedical information together with psychological, lifestyle, and social factors, providing holistic assessment and care, especially for complex or chronic conditions.

B PARADIGM-SPECIFIC KEYWORDS

To support automated paradigm localization and adherence evaluation in ParadigmCoT, we curate paradigm-specific keywords for both general and medical domains. During reasoning chain analysis, these keywords serve as triggers to identify paradigm-initiation steps and facilitate subsequent semantic adherence checks. Table 6 summarizes the representative keywords used for each paradigm.

Table 6: Paradigm-triggering keywords are used to automatically identify the initiation of specific reasoning paradigms within the LLM's output.

Paradigm	**Paradigm-Triggering Keywords**
General Paradigms	
Attribution Reasoning	attribution reasoning, attribution, attributional reasoning, because, due to, owing to, attribution, attributed to, evidence suggests, substantiated by, proven by, grounded in, based on, according to, cross-check, verifiable, claim is supported, factual basis, granular evidence
Counterfactual Reasoning	counterfactual reasoning, counterfactual, counterfactual analysis, what if, if it were not for, alternatively, otherwise, suppose that, imagine if, had it been, contrary to fact, reverse the assumption, changing the input, hypothetically, if we modify, assuming the opposite
Causal Mediation Reasoning	causal mediation reasoning, causal mediation, mediation, mediator, mediated by, intermediate variable, intermediary, intervene, intervention, causal link, bridge the gap, regulate, adjustment, process flow, causal pathway, mechanism, through the influence of
Clinical Paradigms in Medicine	
Evidence-Based Medicine	evidence-based medicine, evidence-based, EBM, guideline-directed, meta-analysis, systematic review, best practice, clinical recommendation, high-quality evidence, grading of recommendations, consensus statement, published guideline, strength of evidence, clinical trial, clinical guideline
Precision Medicine	precision medicine, personalized medicine, precision, individualized, patient-specific, tailored therapy, based on genomics, molecular profiling, stratified, risk score, biomarker, comorbidities, genetic testing, personalized intervention, omics-driven, age-adjusted, phenotype-guided, genetic predisposition, actionable mutation, subpopulation, pharmacogenomics
Integrative Medicine	integrative medicine, integrative, holistic medicine, holistic, mind-body, lifestyle modification, dietary intervention, psychosocial, multidisciplinary, patient-centered, wellness, stress reduction, behavioral change, counseling, alternative medicine, social support, functional status, chronic care, health promotion, quality of life, integrative approach, overall well-being

C CUSTOM MEDICAL BENCHMARKS

To further evaluate clinical reasoning and real-world medical decision making, we employ two carefully curated in-house medical benchmarks. Both benchmarks were independently annotated in a double-blind manner by two physicians with over eight years of clinical experience. The initial inter-annotator agreement was approximately 90–91%. Following discussion, the consensus rate reached approximately 94–95%; cases without consensus were excluded. Both datasets are balanced by design, with equal numbers of positive and negative labels.

DiagTreat: DiagTreat consists of 100 cases designed to assess the consistency between provided diagnoses and therapeutic interventions in real-world medical scenarios. Each entry includes a clinical context, diagnosis, treatment orders, a binary label indicating whether the treatments are appropriate and consistent with the diagnosis, and an explanatory rationale. The benchmark covers 20 common diagnoses and contains balanced Yes/No labels (1:1).

Table 7: Representative DiagTreat Sample

Field	**Content**
Clinical Scenario	Patient fell 4 days ago, resulting in pain and swelling in the left elbow. No loss of consciousness, normal bowel/bladder. Emergency department X-ray: left olecranon and coronoid process fracture; possible avulsion fracture of distal humerus. Cast immobilization started. Follow-up CT: clear displacement of coronoid process and widening of the humero-ulnar joint space. MRI re-check recommended as needed.
Diagnosis	Left elbow fracture
Treatment Orders	Cast fixation; repeat imaging and MRI as indicated.
Consistency Label	No
Explanation	The post-admission management lacks definitive treatment for the displaced elbow fracture, such as surgical intervention or anatomical reduction, and thus does not fully address the diagnosis.

MedJudge: MedJudge contains 98 records for evaluating the appropriateness of drug prescriptions given patient presentations, laboratory results, and comorbidities. Each case provides a clinical narrative, relevant laboratory and examination findings, prescribed medications, a binary Yes/No appropriateness label, and an explanatory rationale. The benchmark covers 31 common medications and has a balanced Yes/No label distribution (1:1).

Table 8: Representative MedJudge Sample

Field	**Content**
Clinical Case	Elderly male, admitted for chronic anemia. No dizziness or palpitations. Diabetes and hypertension history. Labs: Hb 104g/L, Hct 36.1%, WBC 11.29×10^9/L, PLT 180×10^9/L.
Prescription	Oral iron supplements and red blood cell transfusion.
Appropriateness Label	No
Explanation	Symptoms from low hemoglobin are not prominent; iron therapy is not indicated until the etiology of the anemia is clarified.

D PROMPT TEMPLATES

Table 9: Prompt for base and instruct models on the GSM8K and MATH500 datasets.

Model Type	Prompt
Base model	A conversation between User and Assistant. The user asks a question, and the assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. The Assistant shows the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> 12 </answer>. \n\n User: {question} \n Assistant: Let me solve this step by step. <think>
Instruct model	You are a helpful assistant. Given a question, you need to first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. You must show the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> 12 </answer>.

Table 10: Prompt for base and instruct models on the TruthfulQA and HaluEval-QA datasets.

Model Type	Prompt
Base model	A conversation between User and Assistant. The user asks a question, and the assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. The Assistant shows the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> Kim Marton </answer>.\n\n User: {question} \n Assistant: Let me solve this step by step. <think>
Instruct model	You are a helpful assistant. Given a question, you need to first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. You must show the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> Kim Marton </answer>.

Table 11: Prompt for base and instruct models on the HaluEval-QA dataset.

Model Type	**Prompt**
Base model	A conversation between User and Assistant. The user asks a question, and the assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. Now the user asks a question and provides two candidate answers, the assistant needs to determine which answer is correct. The Assistant shows the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> Kim Marton </answer>. \n\n User: #Question#: {question} \n #Candidate Answer 1#: {candidate_answer_1} \n #Candidate Answer 2#: {candidate_answer_2} \n Assistant: Let me solve this step by step. <think>
Instruct model	You are a helpful assistant. Given a question, you need to first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. Now the user asks a question and provides two candidate answers, you need to determine which answer is correct. You must show the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags, such as <answer> Kim Marton </answer>.

Table 12: Prompt for base and instruct models on the MedQA and MMLU-Med datasets.

Model Type	**Prompt**
Base model	A conversation between User and Assistant. The user asks a medical question, and the assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. Now the user asks a question and provides multiple candidate options (e.g., A, B, C, D), and the assistant needs to determine which answer is correct. The Assistant shows the reasoning process within <think></think> tags, and ONLY returns the FINAL ANSWER (option letter) within <answer></answer> tags. For example: <answer> C </answer>. \n\n User: #Question#: question \n #Candidate Option A: option_A \n #Candidate Option B: option_B \n #Candidate Option C: option_C \n #Candidate Option D: option_D \n Assistant: Let me solve this step by step. <think>
Instruct model	You are a helpful assistant. Given a question, you need to first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. You must show the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> C </answer>.

Table 13: Prompt for base and instruct models on the DiagTreat and MedJudge datasets.

Model Type	**Prompt**
Base model	A conversation between User and Assistant. The user asks a question, and the assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. The Assistant shows the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> Yes </answer>.\n\n User: {question} \n Assistant: Let me solve this step by step. <think>
Instruct model	You are a helpful assistant. Given a question, you need to first think about the reasoning process in the mind and then provide the final answer. The reasoning process and answer are enclosed within <think></think> and <answer></answer> tags, respectively. For example, <think> reasoning process here </think> <answer> answer here </answer>. You must show the reasoning process within <think></think> tags, and ONLY return the FINAL ANSWER within <answer></answer> tags. For example: <answer> Yes </answer>.

E SUPPLEMENTARY RESULTS: A CASE STUDY

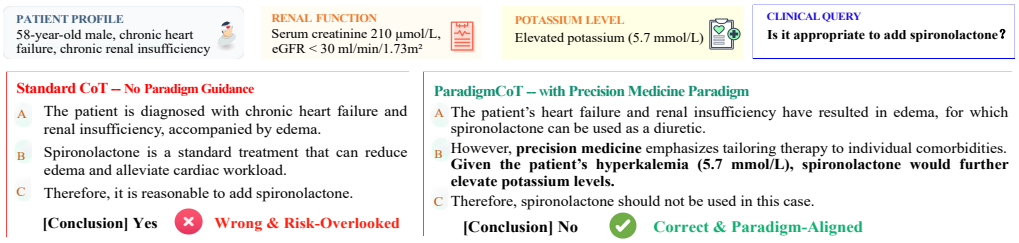


Figure 7: Case study showing how paradigm guidance mitigates constraint drift and enables correct reasoning.

Fig. 7 left illustrates a constraint-drift failure in conventional CoT reasoning. Although the LLM identifies chronic heart failure and partially mentions renal insufficiency, its reasoning shifts toward the generic association that spironolactone is a standard heart-failure treatment, while overlooking the decisive contraindications of elevated potassium and impaired renal function. On the contrary, ParadigmCoT mitigates this drift via the precision medicine paradigm, which emphasizes tailoring decisions to individual comorbidities and risk factors. As a result, ParadigmCoT focuses on the key patient-specific contraindications—hyperkalemia and severely impaired renal function—and reaches the correct conclusion that spironolactone should not be used. This case shows that paradigm guidance makes LLM-generated CoT reasoning more clinically grounded and domain-aligned.