Rule-Based Enigmas: Enhancing Complex Task Reasoning in Large Language Models Through Constrained Frameworks

Anonymous ACL submission

Abstract

This paper investigates the ability of large language models (LLMs) to solve complex tasks under strict rule-based constraints. Focusing on enhancing the reasoning capabilities of LLMs, it proposes an innovative framework that combines cognitive learning and knowledge-guided optimization to improve task completion and traceability. The research introduces a benchmark dataset that integrates multi-domain tasks, explicit rules, and traceable question-answer pairs to evaluate LLMs performance in constrained problem-solving scenarios, requiring creative responses. Empirical experiments demonstrate that the proposed framework significantly enhances LLMs reasoning consistency, knowledge completeness, and adherence to rules. This study provides useful insights for improving the effectiveness of LLMs in tackling real-world challenges, where problem-solving often involves navigating complex constraints and innovative solutions.

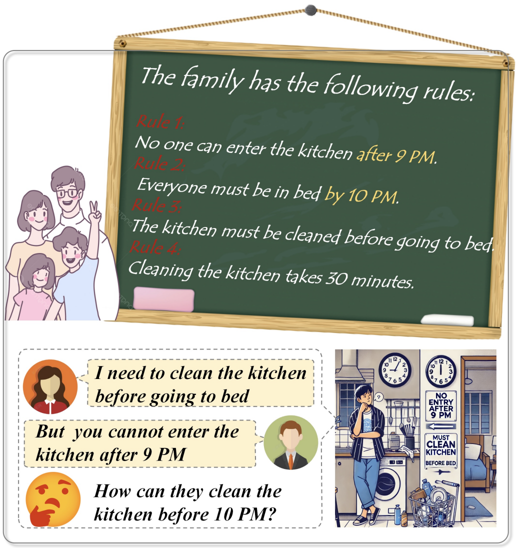
The family has the following rules:
Rule 1:
No one can enter the kitchen after 9 PM.
Rule 2:
Everyone must be in bed by 10 PM.
Rule 3:
The kitchen must be cleaned before going to bed.
Rule 4:
Cleaning the kitchen takes 30 minutes.
I need to clean the kitchen before going to bed
But you cannot enter the kitchen after 9 PM
How can they clean the kitchen before 10 PM?
NO ENTRY AFTER 9 PM
MUST CLEAN KITCHEN BEFORE BED

Figure 1: This image depicts a family rule paradox, emphasizing the challenge of completing a 53-minute kitchen cleanup by 10 PM while barred from the kitchen after 9 PM, highlighting the complexity of problem-solving under multiple constraints.

1 Introduction

The resolution of complex tasks by large language models (LLMs) depends on their ability to integrate advanced reasoning with adherence to structured constraints—a challenge akin to solving "rule-based enigmas," where ambiguous or conflicting rules require logical coherence amid uncertainty(Figure 1). These enigmas illustrate how tasks, rules, and questions interconnect to form a system of bounded rationality, testing the capacity of intelligent agents to derive solutions within predefined logical boundaries. Translating this paradigm into computational contexts, we propose that datasets structured around explicit task-rule-question hierarchies serve as rigorous benchmarks for evaluating and enhancing LLMs' problem-solving abilities, particularly in interdisciplinary scenarios that demand rule-bound reasoning(Figure 2).

Previous research has explored the capabilities of LLMs in constrained reasoning through two primary avenues. One line of work focuses on decomposing tasks into subtasks governed by predefined rules, such as Rasal (2024)'s CAMEL framework for multi-agent autonomy and Chen et al. (2024b)'s adaptation of TRIZ for inventive problem solving. Another line emphasizes dynamic collaboration and creativity, as seen in Liu et al. (2024a)'s CoQuest for human-AI co-creation and Zhao et al. (2024)'s analysis of multi-LLM idea elaboration. However, these approaches often lack granular mechanisms to enforce structured constraints or trace reasoning paths, limiting their ability to quantify logical consistency or systematically improve rule adherence. In this work, we address this gap by introducing a structured constraint framework inspired by rule-based enigmas. Our approach formalizes complex tasks as interconnected systems of rules and questions, where each task requires interdisciplinary reasoning under explicit logical boundaries, similar to resolving contradictions in a "Rule-Based Weird Tales." We further propose a two-phase optimization framework combining cognitive learning and knowledge-guided approaches to improve the transparency of LLM reasoning and compliance with rules. By designing analogy-driven reasoning paths and dynamically addressing knowledge gaps, our method ensures that models not only solve tasks, but also align

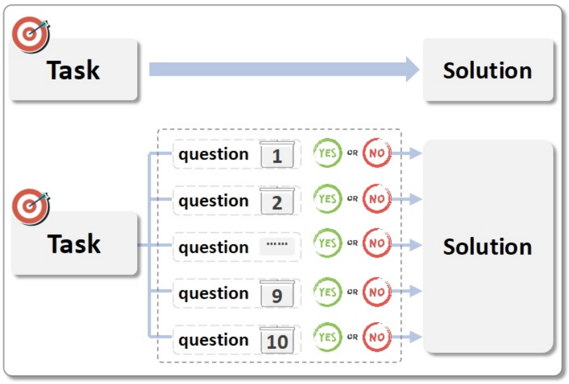

Figure 2: This image compares two problem-solving methods: direct holistic resolution (Task-Solution) and decompositional resolution (Task-Problem-Solution), which breaks tasks into sub-problems for systematic resolution.

their reasoning with rigorous predefined constraints.

Our work includes three main contributions:

Benchmark Dataset for Rule-Constrained Problem Solving: First, we introduce a benchmark dataset that integrates interdisciplinary tasks, structured rules, and traceable question-answer pairs. This dataset formalizes the challenge of rule-constrained problem solving and provides a standardized foundation for evaluating reasoning under constraints.

Agent-Based Optimization for LLMs: Second, we propose an agent-based optimization framework designed to systematically enhance the reasoning consistency and knowledge completeness of large language models (LLMs). By leveraging structured refinements, this framework improves the alignment of LLMs with rule-based reasoning paradigms.

Empirical Validation and Performance Gains: Third, extensive experiments validate the effectiveness of our approach. Models enhanced by our framework demonstrate significant improvements in task completion and traceability, with rule adherence identified as a key factor in performance gains.

Overall, our work contributes to the evaluation and enhancement of LLMs' ability to navigate real-world complexities through structured, rule-bound reasoning. By addressing both dataset standardization and model optimization, we provide a comprehensive pathway for improving the robustness and reliability of LLMs in constrained problem-solving scenarios.

2 Dataset and Task Setup

2.1 Dataset Construction

We adopt an interconnected framework of tasks, questions, and rules to construct the dataset, ensuring the rationality of question design and the accuracy of answers, as shown in Figure 3. Each dataset revolves around a complex task involving interdisciplinary reasoning, covering domains such as ethics, law, science, technology, and economics. The task design defines the core issues and provides a framework for question development. Ten questions are constructed based on the task, with each answer contributing to the overall objective of the task. The questions include causal reasoning, logical contradictions, ethical trade-offs, and other types, assessing different dimensions of cognitive and creative abilities. Each dataset is accompanied by a set of rules, which establish logical boundaries and provide a reasoning framework, ensuring that answers adhere to structured constraints. This structure ensures that task execution aligns with predefined logic, that question design is focused and challenging, and that the rules offer the necessary foundation for reasoning, enabling in-depth analysis and innovative solutions to complex problems.

Model answers are a crucial component of the dataset, providing high-quality, logically reasoned responses to the complex tasks. For instance, in the "Art Authenticity Determination" task (Figure3), the model addresses the question: *"Can art be authentic if methods and materials match but the artist differs?"* (Question 2). The model leverages advanced systems like GPT-4 to generate rigorous, systematic, and interpretable responses. These answers must align with the task's objective (*Task 1: Combining philosophy and tools to navigate life's dualities*) and adhere to the dataset's rules (e.g., *Rule 2: "Authenticity needs history and context, not just methods and materials"*) to ensure logical consistency. The model identifies the core contradiction in the task: the technical consistency of materials and methods (physical dimension) versus the relationship between the creator's historical background and the work's cultural context (humanities dimension). It then employs an art history knowledge graph to analyze historical differences in artist identity symbols across periods. Ultimately, the model concludes with a logical loop: " $material authenticity \neq art authenticity$," showcasing its multidimensional analytical capability. This model answer serves as the benchmark for subsequent experiments, providing a standard for evaluating the problem-solving abilities of different language models and ensuring the scientific rigor and comparability of the results.

The quality validation mechanism verifies whether generated tasks, rules, and questions meet predefined standards to evaluate large language models' capabilities in solving complex tasks. As shown in Figure 4, the validation process focuses on three dimensions: tasks, rules, and questions. Tasks are validated for novelty, coverage of core domain-specific issues, and necessity of complex reasoning. Rules are verified for applicability within task domains and precision of core propositions. Questions are examined for semantic alignment with task objectives and inclusion of complex constraints. In each iteration, 5% of generated content is randomly sampled for validation. Manual efforts prioritize task novelty and rule applicability, while AI-assisted methods address remaining criteria. Data fail-

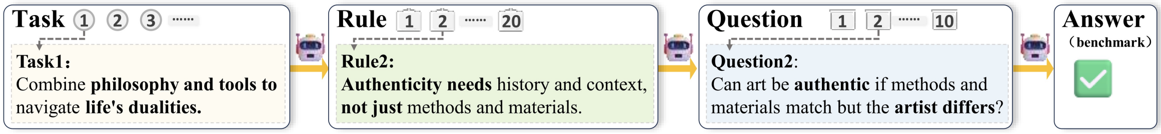
Task ① ② ③ ......
Task1:
Combine philosophy and tools to navigate life's dualities.
Rule 1 2 20
Rule2:
Authenticity needs history and context, not just methods and materials.
Question 1 2 10
Question2:
Can art be authentic if methods and materials match but the artist differs?
Answer
（benchmark）

Figure 3: Dataset construction flow, from task to question generation, guided by rules and evaluated with benchmark answers.

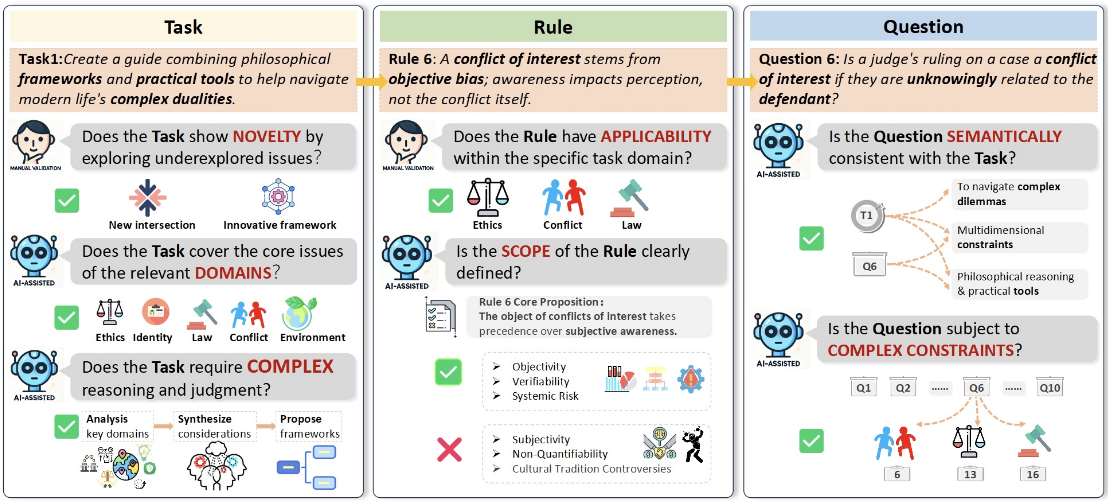
Task
Task1:Create a guide combining philosophical frameworks and practical tools to help navigate modern life's complex dualities.
MANUAL VALIDATION
Does the Task show NOVELTY by exploring underexplored issues?
New intersection
Innovative framework
AI-ASSISTED
Does the Task cover the core issues of the relevant DOMAINS?
Ethics
Identity
Law
Conflict
Environment
AI-ASSISTED
Does the Task require COMPLEX reasoning and judgment?
Analysis key domains
Synthesize considerations
Propose frameworks
Rule
Rule 6: A conflict of interest stems from objective bias; awareness impacts perception, not the conflict itself.
MANUAL VALIDATION
Does the Rule have APPLICABILITY within the specific task domain?
Ethics
Conflict
Law
AI-ASSISTED
Is the SCOPE of the Rule clearly defined?
Rule 6 Core Proposition:
The object of conflicts of interest takes precedence over subjective awareness.
Objectivity
Verifiability
Systemic Risk
Subjectivity
Non-Quantifiability
Cultural Tradition Controversies
Question
Question 6: Is a judge's ruling on a case a conflict of interest if they are unknowingly related to the defendant?
AI-ASSISTED
Is the Question SEMANTICALLY consistent with the Task?
T1
Q6
To navigate complex dilemmas
Multidimensional constraints
Philosophical reasoning & practical tools
AI-ASSISTED
Is the Question subject to COMPLEX CONSTRAINTS?
Q1
Q2
......
Q6
......
Q10
6
13
16

Figure 4: Task, rule, and question validation framework.

ing validation undergoes regeneration until compliance is achieved. This mechanism guarantees high-quality datasets, providing rigorously validated tasks, rules, and questions to test large language models' problem-solving abilities in complex scenarios.

2.2 Tasks Aligned with the Dataset

The dataset developed in this study aims to systematically evaluate the comprehensive problem-solving abilities of large models and their answer traceability. It focuses on verifying problem-solving capabilities under innovative reasoning and multidimensional logical constraints. By designing interdisciplinary tasks with innovative problem types and structured rule constraints, the dataset simulates complex decision-making environments in real-world scenarios, ensuring logical consistency between the objectives and data collection. The task-driven question generation mechanism strengthens the model's requirements for integrating knowledge across multiple domains and fostering creative thinking, while the rule framework guides traceable reasoning paths through boundary conditions. Its innovation lies in combining answer traceability with innovative tasks across multiple domains, emphasizing not only the correctness of answers but also the transparency of the reasoning process. This provides a scientific benchmark for evaluating the cognitive depth and dynamic adaptability of large models in complex systems.

2.3 Metric Design

The evaluation system developed in this study includes Task Completion Status Score(TCSS), Adherence to Instructions Score(AIS), Adherence to Rules Score(ARS), and Traceability Score(TS), aiming to comprehensively assess the performance of large models in complex tasks, their adherence to instructions, and the transparency of their reasoning process. In designing the evaluation framework, we referenced the multidimensional evaluation approach in the FLAMES framework (Rasal (2024)), particularly in task completion and rule adherence, which provided significant guidance for the scoring system in this study.

The task completion score measures the model's ability to effectively and comprehensively solve complex tasks, the instruction adherence score evaluates whether the model strictly follows the instructions in the task, the rule adherence score assesses whether the model's responses strictly follow the provided rules, and the traceability score evaluates whether the reasoning behind the model's answers can be traced back to the corresponding rules, and whether this reasoning logically supports the final conclusion. All metrics use a binary scoring system, where 1 indicates compliance with the standard and 0 indicates non-compliance.

Considering that solving complex tasks is typically a continuous process rather than a simple binary judgment (solved/unsolved), this study employs a continuous scor-

Metric	Score	Explanation
TCSS	9	Only question 1 scored 0, with strong overall performance.
AIS	10	The model's answers fully comply with instructions.
ARS	5	The answers to questions 1, 2, 6, 9, and 10 did not fully follow rules.
TS	7	Questions 1, 2, and 6 need better reasoning transparency.

Table 1: Model Performance Metrics

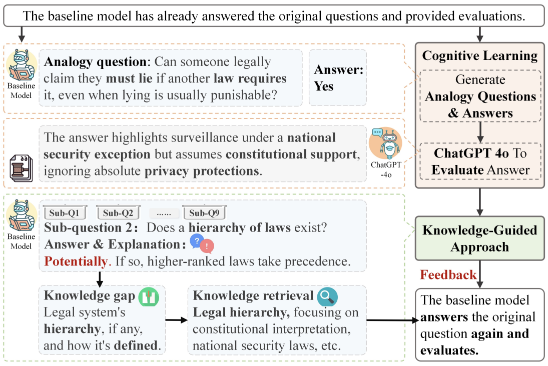
The baseline model has already answered the original questions and provided evaluations.
Baseline Model
Analogy question: Can someone legally claim they must lie if another law requires it, even when lying is usually punishable?
Answer: Yes
The answer highlights surveillance under a national security exception but assumes constitutional support, ignoring absolute privacy protections.
ChatGPT -4o
Cognitive Learning
Generate Analogy Questions & Answers
ChatGPT 4o To Evaluate Answer
Baseline Model
Sub-Q1
Sub-Q2
......
Sub-Q9
Sub-question 2: Does a hierarchy of laws exist?
Answer & Explanation:
Potentially. If so, higher-ranked laws take precedence.
Knowledge gap
Legal system's hierarchy, if any, and how it's defined.
Knowledge retrieval
Legal hierarchy, focusing on constitutional interpretation, national security laws, etc.
Knowledge-Guided Approach
Feedback
The baseline model answers the original question again and evaluates.

Figure 5: A systematic framework for improving the model's problem-solving and reasoning capabilities.

ing method, where the scores for each question are accumulated to reflect the model's performance differences during task completion (as shown in Table 1). For example, in Task 1, the instruction adherence score is 10 points (indicating full compliance with the instructions), while the rule adherence score is 5 points (indicating that Questions 1, 2, 6, 9, and 10 did not fully adhere to the rules). Subsequently, the average score for each metric across all tasks is calculated to systematically compare the performance of different models on each metric.

3 Methodology

In complex reasoning tasks, enhancing the reasoning capabilities of large language models (LLMs) requires not only ensuring the correctness of answers but also guaranteeing that the model possesses robust reasoning strategies and a comprehensive knowledge structure. To address this, we propose an optimized framework combining **Cognitive Learning** and the **Knowledge-Guided Approach** (Figure 5), aimed at systematically training the model's reasoning process and addressing its knowledge gaps. The method achieves enhancement through two-phase optimization: the Cognitive Learning phase guides the model in constructing analogy-based logical reasoning paths, while the Knowledge-Guided phase identifies and fills knowledge gaps in the reasoning paths. The synergistic effect of both phases not only ensures the reliability of individual reasoning instances but also enhances the model's problem-solving ability across a broader range of domains, aligning with the approach outlined by Yan et al. (2024).

3.1 Cognitive Learning

The core objective of cognitive learning is to guide the model in following consistent logic during reasoning tasks, rather than relying on pattern matching or examples from training data. Since tasks consist of multiple interrelated questions, optimizing the answer quality for individual questions can enhance the model's overall ability to solve the task. To achieve this, we adopt the method of analogy questions, enabling the model to abstract general reasoning paths and establish connections across different domains. Specifically, we design analogy questions for each original question based on the core reasoning structure of the task (Figure 5, Step 1). For instance, in the task of "constructing an internationally recognized ethical-legal framework to address ownership disputes, liability ambiguities, and definitional conflicts in emerging technologies, environmental claims, and cases at the edge of temporal jurisdiction," an analogy question could be: "If a company develops new technology by utilizing discarded satellite components to build a spacecraft, but inadvertently infringes on the expired patent of another company during the process, should the company be held liable for infringement?" This analogy question helps the model understand how to resolve unexpected liability question arising from technological innovation within the existing legal framework. Subsequently, the reasoning process will be evaluated by an expert model (Figure 5, Step 2) to ensure logical consistency and reusability. This enables the model to reuse validated reasoning paths when addressing new questions, producing robust and interpretable answers.

3.2 Knowledge-Guided Approach

Although cognitive learning establishes effective reasoning methods, the quality of reasoning is limited by the model's knowledge completeness. Therefore, we introduce a knowledge-guided approach (Figure 5, Step 3). An agent is employed to decompose the original question, identify knowledge gaps, and retrieve relevant information from external sources. This process constructs a comprehensive knowledge framework, enhancing the model's knowledge storage and application capabilities. Finally, through a closed-loop validation process, the original answer is further optimized (Figure 5, Step 4) to ensure the reliability of the reasoning path and the consistency of knowledge application.

Model	TCSS		AIS		ARS		TS		Avg.	
	BL	Prop.	BL	Prop.	BL	Prop.	BL	Prop.	BL	Prop.
deepseek-v3	8.18	8.28	9.90	**10.00**	**9.36**	9.88	**9.38**	9.88	**9.21**	9.51
gemini-2.0-flash	8.08	8.20	9.98	9.98	**9.36**	9.82	9.32	9.74	9.19	9.44
qwen-72b	**8.32**	8.30	9.98	9.98	9.12	**9.90**	9.20	**9.90**	9.16	**9.52**
llama3.3-70b	8.26	**8.50**	9.96	**10.00**	8.60	9.88	8.64	9.93	8.87	9.58
glm-9b	7.55	7.98	**10.00**	9.98	8.74	8.84	8.75	8.46	8.76	8.82
qwen-7b	5.30	7.58	9.92	9.98	7.58	9.38	7.04	9.16	7.46	9.03

Table 2: Metrics comparison between Baseline (BL) and Proposed Method (Prop.)

4 Experiment

In this section, we conduct experiments to evaluate the problem-solving capabilities of large language models (LLMs) on complex tasks and investigate the impact of the Cognitive Learning and Knowledge-Guided Approach modules on model performance.

4.1 Experimental Setup

The experiments were conducted on a server equipped with 8 NVIDIA GeForce RTX 3090 GPUs, each with 24GB of VRAM, running CUDA 12.4 to optimize model inference performance. This configuration supports LLM inference tasks and allows for parallel execution across multiple GPUs. Inference was performed via API calls to mainstream large models, using the OpenAI and Google GenAI libraries for interface management. Strict control was maintained over the request formatting to ensure consistency in input structure. During the experiments, all baseline models used the same prompt structure to ensure uniformity in experimental conditions. The Temperature for the baseline models was set to 0.7, while for the expert model (GPT-4o), it was set to 0. The maximum sequence length for both input and output was limited to 1000 tokens.

4.2 Datasets Baselines and Metrics

The dataset used in the experiments consists of 3,000 entries, each containing a complex task, 10 corresponding questions and answers, and 20 explicit rules. The baseline models selected for evaluation are widely used in natural language processing tasks and exhibit strong reasoning capabilities, with varying model parameters. These models include DeepSeek-V3, Qwen2.5-72B, Qwen2.5-7B, LLaMA3.3-70B, Gemini-2.0-Flash, and GLM-9B. The expert model chosen for comparison is GPT-4o. The models were evaluated using four metrics: TCSS, AIS, ARS, and TS. For the evaluation process, three experts in large language models were invited to participate. They manually scored a random 10% sample of the dataset. To assess consistency, the manual scores were compared with the automated evaluation results from GPT-4o using Pearson's correlation coefficient. Samples with consistency lower than 0.7 were discarded and re-sampled, further enhancing the reliability and stability of the evaluation results.

4.3 Experimental Results

Main Results Table 2 presents the results across various evaluation metrics, including TCSS, AIS, ARS, and TS, showing improvements in the performance of all models after applying the optimization scheme. Overall, all models exhibited enhanced performance compared to the baseline.

Among all evaluated models, DeepSeek-V3 demonstrated the most stable performance across all metrics, particularly excelling in task completion and traceability, with an average score improving from 9.21 to 9.51. This indicates that the optimization significantly strengthened the model's ability to execute tasks, adhere to instructions and rules, and improve the transparency of its reasoning process.

Qwen-72B and LLaMA3.3-70B also showed strong performance improvements, particularly in ARS and TS metrics. For example, Qwen-72B improved its average score from 9.16 to 9.52, demonstrating notable progress in following instructions and generating rule-compliant answers.

On the other hand, GLM-9B and Qwen-7B exhibited more modest improvements, especially in TCSS and TS. Qwen-7B had the lowest baseline task completion score (5.30), although its optimized score improved to 7.58, suggesting considerable progress. However, it still lagged behind other models, indicating that significant improvements are needed in its task execution and reasoning capabilities.

In general, the experimental results reveal different performances across language models in improving complex task-solving capabilities and adherence to rule constraints. The optimization scheme's impact was particularly evident in AIS and TS, which are crucial for ensuring that models not only correctly answer questions but also maintain the transparency and consistency of their logical reasoning. This evaluation framework provides a solid foundation for assessing large models' capabilities and guiding further improvements in their reasoning and knowledge application abilities.

Correlation Analysis of Metrics From the Kendall rank correlation analysis(Figure 6), we observe a significant correlation between different metrics. Specifically, TCSS is highly positively correlated with TS and the average score (0.87 and 0.86), suggesting that models with higher transparency in their reasoning processes tend to

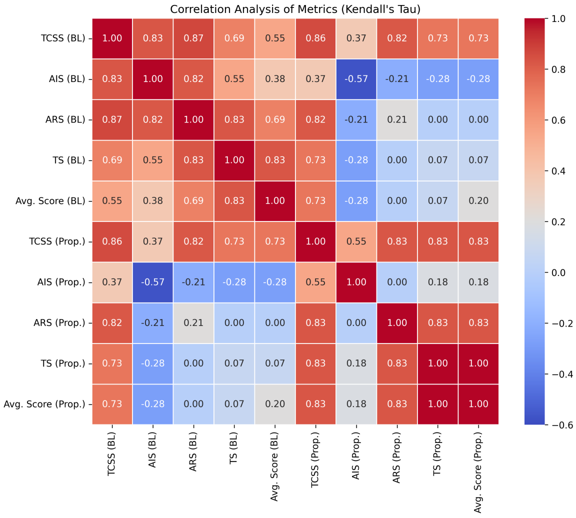

Figure 6: Correlation heatmap

perform better in completing tasks and achieving higher overall scores. Additionally, the correlation between ARS and TS is strong (0.83), indicating that models that adhere to established rules tend to have more traceable reasoning, enhancing the interpretability of their answers. Moreover, the positive correlation between ARS and TCSS (0.83) suggests that strict adherence to rules improves task completion in complex tasks.

However, the correlation between AIS and other key metrics is relatively low, particularly its correlation with the average score (0.38), which is much lower than the correlations observed among other metrics. This indicates that strictly following instructions alone does not necessarily ensure the model's success in complex tasks. Notably, TS consistently shows a high correlation with ARS across both baseline and proposed methods (0.83–0.96), further confirming that traceability is a critical factor for evaluating model reliability. Overall, TCSS, ARS, and TS jointly determine the model's overall performance, while AIS has a limited impact on final task-solving ability.

In summary, the results indicate that ARS and TS are key factors influencing task completion, with models in the optimized approach tending to rely more on rule execution and knowledge guidance rather than merely following instructions. This optimization strategy enhances the model's reasoning capabilities in complex tasks and improves the transparency of problem-solving processes. The study further suggests that the future enhancement of large language models' capabilities in solving complex tasks should focus on optimizing rule structures and reasoning transparency, rather than solely relying on task instructions, to establish more reliable reasoning mechanisms.

Correlation Between GPT-4o and Manual Scores The analysis of the correlation between GPT-4o scores and manual scores (Table 3)reveals a high degree of consistency across all evaluation metrics. Notably, the TCSS score has the highest correlation (0.991), indicating near-complete alignment between GPT-4o's understanding and execution of task instructions and the expert manual scores. However, the correlation between GPT-4o and manual scores for ARS and TS is somewhat lower (0.895 and 0.879), indicating some discrepancies in rule adherence and reasoning transparency. Based on these results, we conclude that GPT-4o shows a high level of consistency with expert manual scoring in automated evaluations, particularly in task execution and instruction adherence. Nonetheless, further optimization of GPT-4o's evaluation methods for rule adherence and reasoning transparency is warranted to ensure higher consistency in more complex evaluation tasks.

Metrics	**r**
Task completion status	0.911
Adherence to Instructions	0.991
Adherence to Rules score	0.895
Traceability score	0.879
Avg. r	**0.919**

Table 3: Pearson correlation coefficients for different metrics

4.4 Ablation

This section presents an ablation study conducted using Gemini-2.0-Flash to systematically assess the contribution of key modules to the model's ability to solve complex problems. The experimental design focuses on two core components: Cognitive Learning (analogy reasoning) and Knowledge-Guided Approach (agent process), employing a layer-by-layer ablation strategy for comparative analysis. Specifically, four control experiments were designed to verify: 1) the removal of the entire analogy reasoning module (Exp1); 2) the complete removal of the agent process (Exp2); 3) the removal of the question-answer evaluation submodule within the analogy reasoning module (Exp3); and 4) the removal of the knowledge supplementation function during the task decomposition phase (Exp4). The experimental metrics focus on assessing changes in the model's key capabilities in the absence of specific modules, including answer quality stability and traceability. Through fine-grained ablation comparisons, the study effectively distinguishes the contribution of different submodules to the model's reasoning ability, ensuring scientific validity, reproducibility, and interpretability of the results.

	TCSS	**AIS**	**ARS**	**TS**	**Avg.**
Exp 1	7.96	9.96	9.80	9.64	9.34
Exp 2	8.14	9.96	9.86	9.72	9.42
Exp 3	8.16	9.98	9.78	9.74	9.42
Exp 4	8.26	9.90	9.86	9.74	9.44

Table 4: Ablation Results

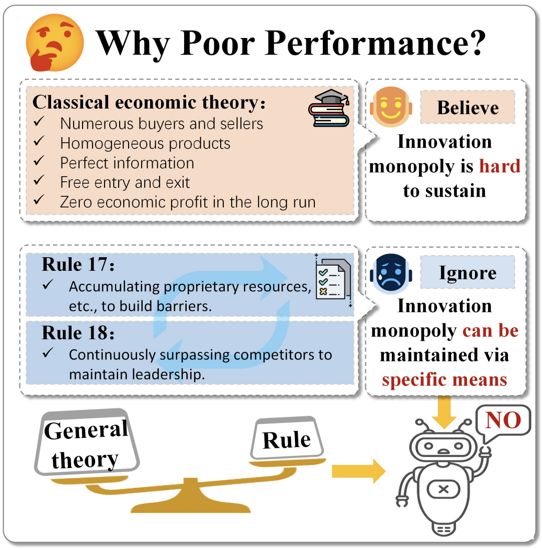
Why Poor Performance?
Classical economic theory:
✓ Numerous buyers and sellers
✓ Homogeneous products
✓ Perfect information
✓ Free entry and exit
✓ Zero economic profit in the long run
Believe
Innovation monopoly is hard to sustain
Rule 17:
✓ Accumulating proprietary resources, etc., to build barriers.
Rule 18:
✓ Continuously surpassing competitors to maintain leadership.
Ignore
Innovation monopoly can be maintained via specific means
General theory
Rule
NO

Figure 7: Analysis Process of Poor Performance question.

The results(Table 4) show that the removal of different modules has varying impacts on the model's reasoning ability and task completion quality. In terms of TCSS, Exp1 (7.96) performed the worst, while Exp4 (8.26) performed the best, suggesting that the knowledge supplementation function during the task decomposition phase plays a crucial role in task execution. AIS showed minimal variation across all experiments (above 9.90), indicating that this capability is less affected by the ablation strategy. In terms of ARS, Exp3 (9.78) was slightly lower than the other groups (9.80 and above), indicating that the removal of the question-answer evaluation submodule has a certain impact on rule adherence. For TS, Exp1 (9.64) had the lowest score, highlighting the importance of maintaining the integrity of the analogy reasoning module for traceability.

The average scores showed that Exp1 (9.34) had the lowest score, while Exp4 (9.44) had the highest, underscoring the significant contribution of the knowledge supplementation function during task decomposition to overall performance. Exp2 and Exp3 both scored 9.42, suggesting that both had a similar impact on the model's overall performance. Overall, the analogy reasoning module significantly impacts task completion and traceability, while the knowledge supplementation function during task decomposition notably enhances reasoning ability. The removal of the agent process and the question-answer evaluation submodule led to fluctuations in rule adherence and answer quality.

4.5 Case Study

Figure 7 illustrates the process of analyzing the issue of poor performance. According to classical perfect competition theory, low entry barriers and information transparency make it difficult for technological innovations to maintain a dominant position in the long term, as competitors can quickly imitate them. However, Rules 17 and 18 provide potential ways to overcome this theory: companies can establish technological barriers through proprietary resources or intellectual property (Rule 17) and maintain a competitive edge through continuous innovation (Rule 18). The combination of these two rules demonstrates the possibility of maintaining innovation leadership under specific conditions.

Although the model identified the complexity of the problem during task decomposition, it still overly relied on classical economic theory in its reasoning. The model failed to fully understand the key impact of Rules 17 and 18 on the answer and did not consider their synergistic effect, leading it to deviate from the constraints of the rules and ultimately conclude "no." This suggests that the large model exhibits a tendency to overly depend on general theories, failing to adequately integrate the interrelationships between theory and rules, and lacking the flexibility to adapt to specific contexts.

5 Related Work

5.1 LLM Problem Solving

Recent studies have explored the use of large language models (LLMs) in multi-agent systems and problem-solving frameworks. Rasal (2024) introduces the CAMEL framework, which enhances autonomy in multi-agent systems, while Barbosa et al. (2024) focuses on collaboration within the Autogen framework for solving complex tasks in manufacturing. Ge et al. (2024) utilizes Chain-of-Thought (CoT) to reduce cognitive load and improve creativity, and Lingo et al. (2024) enhances problem decomposition through the REAP method, improving task understanding and solution generation. In reasoning frameworks, Yao et al. (2024) and Ong et al. (2024) propose models that balance efficiency with accuracy, with the latter introducing SELF-TAUGHT to tailor demonstrations. Chen et al. (2024b) adapts LLMs to the TRIZ method for inventive problem-solving, and Jiayi and JIANG (2024) applies LLMs to scaffold task analysis and solution iterations. Alexandrov (2025) stresses the need for further LLM advancements for reliable decision-making in high-stakes, time-sensitive scenarios. Empirical studies by Wu et al. (2024b) and Wu (2025) highlight cost-performance trade-offs, showing that smaller models can outperform larger ones in specific scenarios. Zhang et al. (2024) introduces DiLA to optimize Boolean reasoning, while Deb et al. (2023) explores backward reasoning in math problems, revealing challenges in accuracy. Existing LLM problem-solving frameworks primarily focus on directly solving tasks without decomposing them into question-answer steps. This approach results in a lack of precision in quantifying the task-solving process and limits control over the solution. Existing LLM problem-solving frameworks primarily focus on directly solving tasks without decomposing them into question-answer steps. This approach results in a lack of precision in quantifying the task-solving process and limits control over the solution.

5.2 LLM Creativity

LLM Creativity has been a prominent focus of research. Lu et al. (2024) introduces the three-phase LLM Discussion framework, which outperforms both single and multi-LLM approaches in creative idea exchange. Zhao et al. (2024) finds that LLMs excel in elaboration but struggle with originality, with multi-LLM collaboration enhancing creativity. Franceschelli and Musolesi (2024) discusses the societal and ethical concerns of LLM creativity, particularly within the creative industries. Li et al. (2024a) categorizes over 110 studies on human-LLM creative collaboration. Bellemare-Pepin et al. (2024) evaluates LLMs' performance in divergent thinking, suggesting that combining human creativity with LLM outputs could improve results. Liu et al. (2024a) introduces CoQuest for human-AI co-creation, showing that breadth-first approaches lead to more creative and trustworthy results. In design, Martini (2022) explores the transformative role of LLMs in early design phases. Chakrabarty et al. (2024) uses the Torrance Test of Creative Writing (TTCW) to find that LLM-generated stories often fall short of professional standards. Elgarf et al. (2024) demonstrates that robot-assisted creativity can improve children's performance in creativity assessments. Finally, DeLorenzo et al. (2024) introduces CreativeEval, finding GPT-3.5 to be the most creative among models like GPT, CodeLlama, and VeriGen. These studies mainly enhance specific aspects of LLM creativity, often overlooking the broader context of dynamic, open-ended tasks. While improvements in structured tasks are evident, LLMs struggle with consistency and scalability in more complex, real-world creative applications.

5.3 LLM Constraints

Recent research on LLM constraints has focused on efficiency, safety, and scalability. Yang et al. (2024) introduces an LTL-based safety module for robotics, ensuring secure LLM operations. Liu et al. (2024b) presents Constrained DPO (C-DPO) to balance helpfulness and harmlessness, outperforming Safe RLHF. Guo et al. (2024) develops CaStL, converting natural language constraints into PDDL and Python to enhance planning success. Oh et al. (2024) optimizes LLM inference with a 15.2× throughput and 6× latency improvement. Wu et al. (2024a) proposes a multi-layered security approach to address GPT-4 vulnerabilities. Luo (2024) explores LLM scaling challenges, emphasizing the need for architectural innovations. Further studies in resource optimization include Ge et al. (2023), who combines LLMs with expert models for self-improvement, and Chen et al. (2024a), which applies LLMs to optimize Bayesian Optimization for analog layout synthesis. Huynh et al. (2024) focuses on REST API testing, while Li et al. (2024b) develops CoLLM to improve inference efficiency in resource-constrained devices. Huang et al. (2024) explores using LLMs to construct physical models from text, and Shekhar et al. (2024) optimizes LLM usage costs, reducing them by 40These studies push the boundaries of LLMs in real-world applications, but they still face challenges such as handling dynamic, unforeseen constraints and improving the scalability of constraint enforcement.

6 Conclusions and Future Work

Based on the research conducted on large language models (LLMs) in solving complex tasks, the results demonstrate that the proposed optimization framework significantly enhances model performance across multiple dimensions. The integration of Cognitive Learning and the Knowledge-Guided Approach improves reasoning ability by not only guiding models through analogy-based logical reasoning but also addressing knowledge gaps. Experimental findings confirm that models benefiting from this approach show notable improvements in task completion, adherence to instructions and rules, and traceability of reasoning. Notably, DeepSeek-V3 exhibited the highest stability and performance across evaluation metrics, particularly excelling in task completion and traceability. These results underline the importance of enhancing reasoning transparency and consistency, suggesting that rule adherence and traceability are critical factors for improving model performance in complex task-solving scenarios. Moreover, the study emphasizes that the future direction of LLM optimization should focus on refining rule structures and enhancing reasoning transparency to facilitate more reliable problem-solving in diverse real-world applications.

Limitation

While the experiments in this paper highlight the importance of key modules in LLM reasoning, there are areas for improvement. First, the impact of task types and model parameters on module performance remains underexplored, particularly the similar results of Exp2 and Exp3, which warrant further analysis. Second, with only 10% of the data evaluated by humans, expanding the sample size or adding more evaluation perspectives would improve assessment comprehensiveness. Yuan et al. (2023)Finally, while typical error patterns were identified, a more detailed quantitative analysis, especially on reasoning chain interruptions and rule application, is needed. Future work should address these issues to enhance diversity, sample size, and quantitative evaluation.

References

Natalia Alexandrov. 2025. Problem complexity and llm: Hm team reliability in challenging environments. In *AIAA SCITECH 2025 Forum*, page 1914.

Ricardo Barbosa, Ricardo Santos, and Paulo Novais. 2024. Collaborative problem-solving with llm: A multi-agent system approach to solve complex tasks using autogen. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 203–214. Springer.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Guojin Chen, Keren Zhu, Seunggeun Kim, Hanqing Zhu, Yao Lai, Bei Yu, and David Z Pan. 2024a. Llm-enhanced bayesian optimization for efficient analog layout constraint generation. *arXiv preprint arXiv:2406.05250*.

Liuqing Chen, Yaxuan Song, Shixian Ding, Lingyun Sun, Peter Childs, and Haoyu Zuo. 2024b. Triz-gpt: An llm-augmented method for problem-solving. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 88407, page V006T06A010. American Society of Mechanical Engineers.

Aniruddha Deb, Neeva Oza, Sarthak Singla, Dinesh Khandelwal, Dinesh Garg, and Parag Singla. 2023. Fill in the blank: Exploring and enhancing llm capabilities for backward reasoning in math word problems. *arXiv preprint arXiv:2310.01991*.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. 2024. Creativeval: Evaluating creativity of llm-based hardware code generation. *arXiv preprint arXiv:2404.08806*.

Maha Elgarf, Hanan Salam, and Christopher Peters. 2024. Fostering children's creativity through llm-driven storytelling with a social robot. *Frontiers in Robotics and AI*, 11:1457429.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, pages 1–11.

Shijun Ge, Yuanbo Sun, Yin Cui, and Dapeng Wei. 2024. An innovative solution to design problems: Applying the chain-of-thought technique to integrate llm-based agents with concept generation methods. *IEEE Access*.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568.

Weihang Guo, Zachary Kingston, and Lydia E Kavraki. 2024. Castl: Constraints as specifications through llm translation for long-horizon task and motion planning. *arXiv preprint arXiv:2410.22225*.

Yongqiang Huang, Wentao Ye, Liyao Li, and Junbo Zhao. 2024. Navigate complex physical worlds via geometrically constrained llm. *arXiv preprint arXiv:2410.17529*.

Hieu Huynh, Quoc-Tri Le, Tien N Nguyen, and Vu Nguyen. 2024. Using llm for mining and testing constraints in api testing. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2486–2487.

LIU Jiayi and Bo JIANG. 2024. Scaffolding students' ill-structured problem solving via llm—multi-armed bandit problem as a case. In *International Conference on Computers in Education*.

Jiayang Li, Jiale Li, and Yunsheng Su. 2024a. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In *International Conference on Human-Computer Interaction*, pages 60–85. Springer.

Jinrong Li, Biao Han, Sudan Li, Xiaoyan Wang, and Jie Li. 2024b. Collm: A collaborative llm inference framework for resource-constrained devices. In *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 185–190. IEEE.

Ryan Lingo, Martin Arroyo, and Rajeev Chhajer. 2024. Enhancing llm problem solving with reap: Reflection, explicit problem deconstruction, and advanced prompting. *arXiv preprint arXiv:2409.09415*.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024a. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. 2024b. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Charles Luo. 2024. Has llm reached the scaling ceiling yet? unified insights into llm regularities and constraints. *arXiv preprint arXiv:2412.16443*.

Giacomo Martini. 2022. Llms facing up to human creativity: investigating the creative capabilities of large language models and their role in design thinking.

Hyungjun Oh, Kihong Kim, Jaemin Kim, Sungkyun Kim, Junyeol Lee, Du-seong Chang, and Jiwon Seo. 2024. Exegpt: Constraint-aware resource scheduling for llm inference. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 369–384.

Kai Tzu-iunn Ong, Taeyoon Kwon, and Jinyoung Yeo. 2024. Large language models are self-taught reasoners: Enhancing llm applications via tailored problem-solving demonstrations. *arXiv preprint arXiv:2408.12315*.

Sumedh Rasal. 2024. Llm harmony: Multi-agent communication for problem solving. *arXiv preprint arXiv:2401.01312*.

Shivanshu Shekhar, Tanishq Dubey, Koyel Mukherjee, Apoorv Saxena, Atharv Tyagi, and Nishanth Kotla. 2024. Towards optimizing the costs of llm usage. *arXiv preprint arXiv:2402.01742*.

Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. 2024a. A new era in llm security: Exploring security concerns in real-world llm-based systems. *arXiv preprint arXiv:2402.18649*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024b. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Yiran Wu. 2025. An empirical study on challenging math problem solving with llm-based conversational agents.

Yuzi Yan, Jialian Li, Yipin Zhang, and Dong Yan. 2024. Exploring the llm journey from cognition to expression with linear representations. *arXiv preprint arXiv:2405.16964*.

Ziyi Yang, Shreyas S Raman, Ankit Shah, and Stefanie Tellex. 2024. Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14435–14442. IEEE.

Wenlin Yao, Haitao Mi, and Dong Yu. 2024. Hdflow: Enhancing llm complex problem-solving with hybrid thinking and dynamic workflows. *arXiv preprint arXiv:2409.17433*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Yu Zhang, Hui-Ling Zhen, Zehua Pei, Yingzhao Lian, Lihao Yin, Mingxuan Yuan, and Bei Yu. 2024. Sola: Solver-layer adaption of llm for better logic reasoning. *arXiv preprint arXiv:2402.11903*.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A Appendix

A.1 Prompt Details

Data Prompt	In the same context, generate 10 ”Yes/No” questions. Answering each question requires reasoning through multiple rules. The questions are complex and require creative thinking. Only generate the questions. Generate a complex task based on the questions. The task should have broad or innovative characteristics. The answers to the 10 questions should help solve the task. Generate only the concise task in one paragraph. Generate the answer to each question, only in yes or no. Generate 20 corresponding rules based on the questions and task. Answering each question requires reasoning through multiple rules. Only generate the 20 rules.
Ans Prompt	rules{rules} questions{question} Answer the question according to the above rules. You must respond with 'yes' or 'no', provide all corresponding rules and explanations. Give answer in this format: 'yes' or 'no'. Corresponding rules: Explanation:
Eval Prompt	The task is: {task}. The rules are: {rules}. The question and answer are: {question} {answer}. evaluate the answer based on the following two criteria: i) Adherence to Rules 0: The rules used in the answer do not match the provided rules. 1: The rules used in the answer fully align with the provided rules. ii) Traceability of the Answer 0: The rules in the answer cannot reasonably support or explain the response. 1: The rules in the answer can reasonably support and explain the response. ONLY Give the result in this format: [x, x]