Dissecting Logical Reasoning in LLMs: A Fine-Grained Evaluation and Supervision Study

Yujun Zhou[1,*], **Jiayi Ye**[2,*], **Zipeng Ling**[3,*], **Yufei Han**[4], **Yue Huang**[1], **Haomin Zhuang**[1]
Zhenwen Liang[1] **Kehan Guo**[1], **Taicheng Guo**[1], **Xiangqi Wang**[1], **Xiangliang Zhang**[1]
[1]University of Notre Dame [2]MBZUAI [3]University of Pennsylvania [4]INRIA
[*]Equal Contribution
{yzhou25,xzhang33}@nd.edu

Abstract

Logical reasoning is a core capability for large language models (LLMs), yet existing benchmarks that rely solely on final-answer accuracy fail to capture the quality of the reasoning process. To address this, we introduce FineLogic, a fine-grained evaluation framework that assesses logical reasoning across three dimensions: overall accuracy, stepwise soundness, and representation-level probing. Leveraging this framework, we conduct a comprehensive study on how different supervision formats in fine-tuning shape reasoning abilities. We fine-tune LLMs on four supervision styles—one in natural language and three symbolic variants—and find a key trade-off: natural language supervision excels at generalization to out-of-distribution and long-chain problems, whereas symbolic supervision is superior at instilling structurally sound, atomic reasoning steps. Furthermore, our probing analysis indicates that fine-tuning primarily refines the model's step-by-step generation process, rather than improving its ability to converge on an answer early. Together, our framework and analysis provide a more rigorous lens for evaluating and improving logical reasoning in LLMs. The code is available at https://github.com/YujunZhou/FineLogic.

1 Introduction

Large language models (LLMs) are rapidly emerging as transformative tools across a wide array of applications (Achiam et al., 2023; Guo et al., 2024b; Thirunavukarasu et al., 2023; Nam et al., 2024; Huang et al., 2024, 2025; Zhou et al., 2024a). Among these, reasoning serves as a core capability underpinning tasks such as problem-solving (Lu et al., 2023), scientific question answering (Guo et al., 2024a; Zhou et al., 2024b), and code analysis (Nam et al., 2024). Consequently, a growing body of research has sought to evaluate and enhance the reasoning abilities of LLMs from multiple perspectives (Wei et al., 2022; Guo et al., 2025, 2024a; Liang et al., 2024; Wang et al., 2025b; Zhou et al., 2025; Zhao et al., 2025). Within this broader landscape, logical reasoning stands out as a particularly challenging and intellectually demanding domain (Saparov and He, 2022). It requires a synthesis of natural language understanding, formal logical interpretation, and multi-step inferential processing (Patel et al., 2024; Saparov et al., 2023; Morishita et al., 2024).

Despite growing interest in the logical reasoning capabilities of LLMs, most existing benchmarks focus narrowly on whether a model produces the correct final answer (Patel et al., 2024; Parmar et al., 2024; Han et al., 2022). This binary evaluation, typically assessing only the correctness of a "True" or "False" output, can be misleading, as it fails to determine whether the model arrived at the answer through valid multi-step reasoning (Saparov and He, 2022). Consequently, correct answers may reflect guesswork rather than genuine logical inference. We are thus motivated to address **RQ1: How to rigorously evaluate LLMs' step-by-step correctness in logical reasoning tasks, beyond the binary evaluation of the final answer?**

In parallel with benchmarking efforts, numerous methods have been proposed to enhance the multi-step logical reasoning abilities of LLMs. While many leverage inference-time strategies (Wang et al., 2025a), in-context learning (Creswell et al., 2022; Xu et al., 2024), or external logical verifiers (Pan et al., 2023) to guide the model toward more rigorous reasoning, some recent studies explored supervised fine-tuning (SFT) as a more direct approach to enhancing logical reasoning (Morishita et al., 2024; Feng et al., 2023). For example, Morishita et al. (2024) proposes a synthetic logic corpus designed to offer broad and systematic coverage of logical knowledge. However, it remains unclear for this important question, **RQ2: What style of**

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17075–17098
November 4-9, 2025

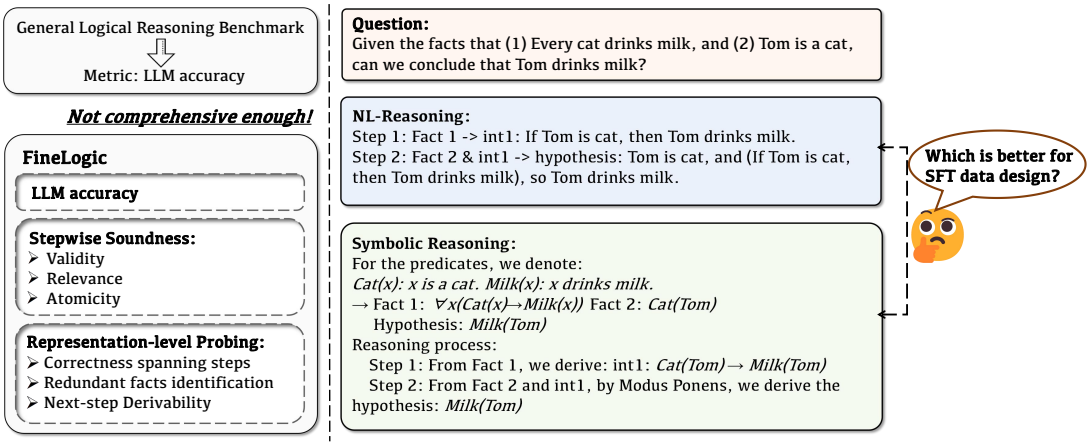


Figure 1: (Left) LLM logical reasoning evaluation: the general benchmark v.s. our fine-grained benchmark **FineLogic**. (Right) processing a logical reasoning task using natural language v.s. using symbolic methods.

training data, natural language or formal logical symbols, better facilitates the learning of multi-step logical reasoning through SFT? Addressing this research question is important for understanding how to most effectively instill logical reasoning capabilities in LLMs.

To address RQ1, we propose **FineLogic**, a diagnostic framework designed to offer a multi-dimensional evaluation of LLMs' reasoning processes, moving beyond binary final-answer correctness. Rather than creating another leaderboard, FineLogic serves as a tool for LLM practitioners to pinpoint specific weaknesses in reasoning chains, such as flawed logic, redundancy, or non-atomic steps. Our framework evaluates models along three complementary dimensions: (1) **Overall benchmark accuracy**: This metric captures a model's ability to perform multi-step logical reasoning and its generalizability across problems from diverse domains. (2) **Stepwise Soundness**: Inspired by Saparov and He (2022), we assess the quality of each intermediate reasoning step using three criteria—**validity** (whether the step is logically valid), **relevance** (whether its conclusion is used in later steps), and **atomicity** (whether it applies a single, minimal inference rule). These metrics aim to evaluate the model's ability to generate human-interpretable and logically coherent reasoning chains. (3) **Representation-level probing** (Ye et al., 2024): By applying probing techniques to LLM hidden representations, this evaluation provides insight into whether the model's understanding of logical structure is merely surface-level or embedded in its internal state.

To address RQ2, we systematically investigate how different supervision formats affect the reasoning capabilities of LLMs. Specifically, we examine both natural language-based training data and logic-symbol-based representations, including several structured variants. Our analysis shows that **natural language supervision is particularly effective in conveying core reasoning patterns**, leading to strong performance across a wide range of evaluation benchmarks. Notably, it exhibits impressive **generalizability even on out-of-distribution test sets** that require long reasoning chains. However, a deeper examination of *stepwise soundness* and *internal representation probing* reveals certain limitations. Models trained with natural language supervision tend to struggle with producing strictly minimal reasoning chains (e.g., more likely including **redundant steps** and applying multiple inference rules in a single step, as shown in Figure 5). In contrast, models trained with symbolic reasoning styles are better at filtering out irrelevant information, generating atomic steps aligned with individual deduction rules, and maintaining cleaner, logically grounded reasoning trajectories.

To summarize, our contributions are as follows:

- We propose FineLogic, a unified diagnostic framework for assessing LLMs' logical reasoning. It moves beyond final-answer accuracy to evaluate the quality and coherence of reasoning steps, offering fine-grained insights that can guide targeted model improvements and serve as a scaffold for structured reward design in reinforcement learning.

- We conduct a comprehensive study on the effects of supervision format, fine-tuning LLMs on both natural language and symbolic logic data to examine their impact on reasoning across general and complex tasks.
- Through systematic analysis of models trained with different supervision styles, we identify key trade-offs between generalization and structural reasoning quality. The findings provide concrete insights into the design and selection of effective training data for post-training.

2 Related Works

Logical Reasoning Benchmarks. Numerous benchmarks have been proposed to evaluate the logical reasoning abilities of LLMs. Many either mix logical and commonsense reasoning (Liu et al., 2023; Luo et al., 2023; Havrilla et al., 2024), making it hard to isolate logical competence, or assess multi-step reasoning using only final-answer accuracy (Parmar et al., 2024; Han et al., 2022; Tafjord et al., 2020; Mondorf and Plank, 2024). While ProntoQA (Saparov and He, 2022; Saparov et al., 2023) pioneered stepwise evaluation, its analysis is confined to step correctness on short problems. FineLogic moves beyond this by introducing a more comprehensive suite of step-level diagnostics—including a more practical relevance metric—to rigorously assess complex, long-chain reasoning. Furthermore, we are the first to adapt representation-level probing from mathematics (Ye et al., 2024) to the logical domain, introducing novel tasks like Correctness Spanning Steps (CSS) to connect behavioral outputs with internal model states. In contrast, other stepwise frameworks like ROSCOE (Golovneva et al., 2022) and RECEVAL (Prasad et al., 2023) remain too generic or coarse-grained for the specific demands of formal logic.

Logical Reasoning Enhancement. Several studies have aimed to improve LLMs' performance on logical reasoning tasks. Some approaches rely on translating inputs into formal logic and using programmable verifiers to solve problems (Olausson et al., 2023; Pan et al., 2023; Yang et al., 2023; Ryu et al., 2024), which bypasses the model's own reasoning process. Others use in-context learning or inference-time strategies to guide output without fundamentally enhancing reasoning ability (Creswell et al., 2022; Wang et al., 2025a; Xu et al., 2024; Sun et al., 2023; Toroghi et al., 2024). While a few works have explored fine-tuning or reinforcement learning to strengthen logical reasoning (Feng et al., 2023; Morishita et al., 2023, 2024; Xie et al., 2025; Yang et al., 2022; Xie et al., 2024; Zheng et al., 2025), they have not examined which types of supervision are most effective for teaching LLMs to reason. In this work, we focus specifically on this open question.

3 FineLogic Evaluation Framework

As illustrated in Figure 2, FineLogic builds on existing benchmarks and evaluates logical reasoning ability from three complementary perspectives: (1) **Overall benchmark accuracy**, which measures whether the model can correctly solve multi-step reasoning tasks; (2) **Stepwise soundness**, which evaluates whether each reasoning step is valid and interpretable; (3) **Representation-level probing**, which assesses whether the model internally captures the problem's reasoning structure beyond surface-level patterns.

3.1 Overall Benchmark Accuracy

Similar to most benchmarks, our overall benchmark accuracy focuses on final-answer correctness. Specifically, we define accuracy (Acc) as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i] \tag{1}$$

where N is the number of test problems, y_i is the gold label, and $\hat{y}_i$ is the model's prediction for problem i. While coarse-grained, it offers a quick and effective way to assess a model's overall reasoning ability and cross-domain generalization. We evaluate on four challenging multi-step reasoning benchmarks, deliberately selected for their focus on peer-reviewed, multi-hop formal logic. The suite includes widely-used datasets for comparability (**FOLIO** (Han et al., 2022) and **ProntoQA** (Saparov and He, 2022)) alongside benchmarks designed to probe out-of-distribution generalization by varying complexity and rule composition (**FLD** (Morishita et al., 2024) and **Multi-LogiEval** (Patel et al., 2024)). This selection offers a comprehensive assessment, in contrast to narrower benchmarks like ProofWriter (Tafjord et al., 2020), which we exclude due to its limited scope of logical rules. Further details on dataset statistics and sampling are provided in Table 1 and Appendix A.1.

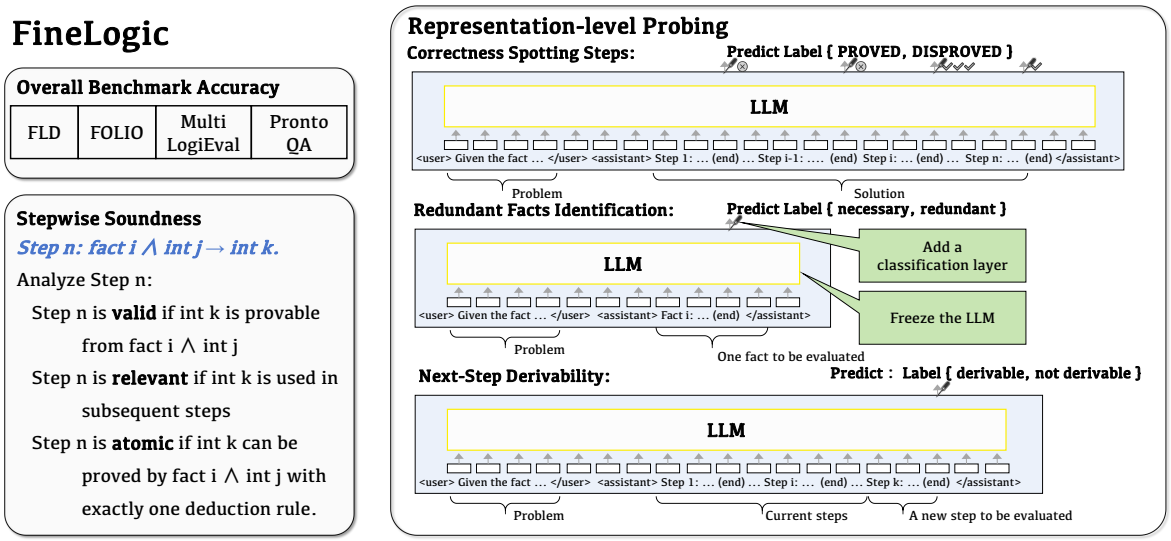


Figure 2: Overview of FineLogic, where overall benchmark accuracy, stepwise soundness, and representation-level probing are combined for a fine-grained evaluation of LLM's logical reasoning ability.

Dataset	Samples	Label Types
FLD (Morishita et al., 2024)	1100	{T, F, Unknown}
FOLIO (Han et al., 2022)	203	{T, F, Unknown}
Multi-Logical (Patel et al., 2024)	390	{T, F}
Pronto-QA (Saparov and He, 2022)	500	{T, F}

Table 1: Sample counts and label types for each dataset.

3.2 Stepwise Soundness

Building on Saparov and He (2022), we evaluate the soundness of each intermediate reasoning step along three dimensions: **validity** (whether the step logically follows from its premises), **relevance** (whether its conclusion is used in later steps), and **atomicity** (whether it applies a single, minimal inference rule).

To assess these criteria, we extract the premises and conclusion of each step from the model's final answer. Crucially, our evaluator scores only the text following a designated answer tag (e.g., </think>), ensuring that any preceding self-reflection or speculative reasoning is excluded by design. We use GPT-4.1-mini to evaluate *validity* and *atomicity*. Manual verification on 200 annotated steps shows that GPT-4.1-mini achieves over 98% accuracy on both metrics. For *relevance*, we determine whether the conclusion of step i is referenced in any subsequent step $k > i$.

We then compute the proportion of samples in which *all* steps are valid, relevant, and atomic, providing a sample-level measure of reasoning integrity. Formally, for each solution s with K_s steps, let $v_{s,k}, r_{s,k}, a_{s,k} \in \{0, 1\}$ denote the validity, relevance, and atomicity of step k, respectively. We define the sample-level metrics as:

$$\text{AllValid} = \frac{1}{N}\sum_{s=1}^{N}\left[\prod_{k=1}^{K_s} v_{s,k}\right] \quad (2)$$

$$\text{AllRelevant} = \frac{1}{N}\sum_{s=1}^{N}\left[\prod_{k=1}^{K_s} r_{s,k}\right] \quad (3)$$

$$\text{AllAtomic} = \frac{1}{N}\sum_{s=1}^{N}\left[\prod_{k=1}^{K_s} a_{s,k}\right] \quad (4)$$

Full prompt templates are provided in Figures 13 and 14.

3.3 Representation-level Probing

Inspired by Ye et al. (2024), we introduce **representation-level probing accuracy** to assess whether LLMs internally understand how and when to perform specific reasoning step. Unlike behavioral metrics, this method aligns internal representations with reasoning structure and tracks how reasoning knowledge evolves across steps.

We construct probing datasets from FLD test samples requiring 10–20 reasoning steps, using 450 problems for training and 100 for testing across three tasks, implementation details are provided in Appendix B:

Correctness Spanning Steps (CSS): Identifies the earliest step after which the model consistently predicts the correct label. The spanning length is the number of remaining steps from that point to the

end. Higher accuracy indicates earlier internalization of the correct answer. Formally, for each problem s, let τ_s be the first step from which the probe consistently predicts the correct label and K_s be the total number of steps. The CSS score is then:

$$\text{CSS} = \frac{1}{N}\sum_{s=1}^{N}(K_s - \tau_s). \tag{5}$$

Redundant Facts Identification (RFI): After presenting all facts and the hypothesis, we append three necessary and three redundant facts. A classifier is trained to distinguish between them, measuring the model's ability to identify irrelevant information. Higher accuracy reflects better fact discrimination.

Next-Step Derivability (NSD): At six randomly selected intermediate steps, we append three valid and three invalid candidate steps. Probing predicts which are currently derivable. Higher accuracy indicates stronger awareness of valid next steps.

For both RFI and NSD, performance is measured using balanced accuracy to account for the construction of positive and negative instances. The score is calculated as:

$$\text{Score} = \frac{1}{2}(\text{TPR} + \text{TNR}),$$

where TPR is the True Positive Rate and TNR is the True Negative Rate.

Our evaluation builds on two prior lines of work—stepwise reasoning evaluation (Saparov and He, 2022) and representation-level probing (Ye et al., 2024)—but introduces key extensions tailored to logical reasoning.

4 Supervision Format and Style: SFT Data Design

In this section, we examine how different supervision styles for SFT affect the logical reasoning abilities of LLMs. Our training data is based on **FLD** and **ProntoQA**, both of which include gold reasoning chains suitable for constructing diverse supervision styles.

For **FLD**, we generate 500 problems for each reasoning depth from 0 to 15, plus 1500 `UNKNOWN` samples, totaling 9500 training instances. For **ProntoQA**, we use 3200 3-hop problems. During evaluation, FLD covers depths 0–19, while ProntoQA uses only the hardest 5-hop samples.

We compare four supervision styles across two categories: natural language-based and symbolic reasoning. Each style reflects a different level of abstraction and clarity in reasoning structure.

- **NL-Reasoning**: Solutions are written entirely in natural language, with no intermediate symbolization or abstraction.
- **Symbolic Reasoning (Structured)**: Problems are formalized by defining variables and predicates, translating facts and hypotheses into logical forms, and reasoning step by step using symbolic logic.
- **Symbolic Reasoning (Filtered)**: A simplified variant where only necessary facts are retained, shortening reasoning chains and reducing input complexity.
- **Symbolic Reasoning (Direct)**: Facts are directly expressed in symbolic form without defining variables or predicates, which shortens sequences but may introduce ambiguity.

A small portion of translations, connective phrases, and intermediate steps are generated using GPT-4.1. Prompt examples are shown in Figure 4 (Appendix E).

5 Experiments

5.1 Experimental Setup

Our experiments involve four base models for fine-tuning—**LLaMA-3.1-8B-Instruct**, **Qwen-2.5-7B-Instruct**, **Qwen-2.5-Math-7B-Instruct**, and **Qwen-3-8B**—and two powerful zero-shot baselines, **GPT-4o** and **DeepSeek R1**. The four base models are fine-tuned for 3 epochs at a 1×10^{-6} learning rate, with their original versions also serving as baselines. Representation-level probing is limited to the LLaMA and Qwen models due to computational constraints, and an explicit step-by-step format is enforced for all stepwise evaluations.

We compare SFT models trained with different supervision styles against these baselines:

- **Direct Answering**
- **Chain-of-Thought (CoT)** (Wei et al., 2022)
- **Few-Shot Learning** (Brown et al., 2020)
- **LOGIPT** (Creswell et al., 2022)
- **Selection-Inference** (Creswell et al., 2022)
- **SymbCoT** (Xu et al., 2024)

- **LogicLM** (Pan et al., 2023)

More detailed experimental setups can be found in Appendix A.

5.2 Results

We conducted experiments for analyzing the performance of four models combined with various prompting and fine-tuning settings under the **Fine-Logic Evaluation Framework**. Due to space constraints, the results for Qwen-2.5-Math-7B-Instruct and Qwen-3-8B are presented in Appendix C.

5.2.1 Results on Overall Benchmark Accuracy

As shown in Table 1, we report the **overall benchmark accuracy** across four datasets, as well as the **step-wise accuracy** on the FLD benchmark, stratified by reasoning depth (Figure 3). Our analysis yields several key observations:

CoT and few-shot prompting generally improve performance, but baseline methods do not consistently yield gains. Across the four evaluation datasets, both CoT and few-shot prompting lead to broadly positive improvements, indicating their general effectiveness in enhancing LLM performance on logical reasoning tasks. Notably, **few-shot prompting consistently outperforms CoT**, suggesting that for complex logical tasks, showing the model *how* to think (via exemplars) is more beneficial than simply encouraging it to reason step by step. This may be because logical questions naturally elicit multi-step reasoning under direct prompting, limiting the marginal benefit of CoT. In contrast, few-shot demonstrations provide clearer procedural scaffolding, which appears more effective in guiding the model's reasoning process.

In contrast, baseline prompting methods such as *Logic-LM*, *SymbCoT*, and *Sel-Inf* show inconsistent performance and sometimes underperform even direct prompting. For example, *Logic-LM* performs well on simpler problems but degrades on complex ones, with Qwen's Multi-LogiEval accuracy dropping to 27.1%. *SymbCoT* sometimes improves over *Logic-LM* (e.g., 63.8% on Multi-LogiEval with Qwen) but also shows large drops elsewhere (e.g., 22.6% on FLD, versus 44.6% with direct prompting).

Supervised fine-tuning outperforms inference-time methods, but its effectiveness heavily depends on the supervision style. Compared to inference-time prompting strategies, SFT yields significantly greater improvements in logical reasoning performance. Among all training styles, **natural language-based supervision (SFT-NL) produces the most substantial and consistent gains across datasets and models.**

Notably, even though SFT was conducted using only problems from FLD and ProntoQA with reasoning depths *less than those in the test set*, the resulting models show robust improvements. For example, under the SFT-NL setting, Llama's accuracy on FLD increased from 31.7% (direct prompting) to 67.5% and Qwen improved from 46.6% to 71.0%, approaching the best-performing baseline DeepSeek R1. On ProntoQA, most SFT variants achieve over 90% accuracy. Furthermore, even on out-of-distribution datasets such as FOLIO and Multi-LogiEval, some SFT settings deliver strong generalization. For instance, on Multi-LogiEval, Llama with SFT-NL improved to 71.3%, matching the performance of GPT-4o.

While SFT-NL demonstrates the best overall and most transferable performance, other styles of supervision yield much smaller gains. This may be since LLMs are primarily pretrained on natural language data, making symbolic reasoning—especially when it requires both translation and inference over logic forms—significantly more challenging. Among the symbolic settings, **SFT-Symb-Filter** consistently outperforms other variants. By removing redundant reasoning steps from the symbolic training data, this setting simplifies training and enhances performance. In contrast, **SFT-Symb-Direct**, which skips variable and predicate definitions entirely, performs poorly, likely due to the introduction of ambiguity and the lack of explicit logical structure.

Accuracy declines with reasoning depth, but SFT enables small models to match GPT-4o even on the most challenging out-of-distribution samples. As shown in Figure 3, model accuracy decreases as the required number of reasoning steps increases. Nonetheless, our results show that SFT substantially improves model robustness, even on long-chain, out-of-distribution examples. On in-distribution FLD test problems (0–15 steps), SFT models trained under most styles outperform GPT-4o. For instance, across reasoning depths up to 15, both Llama and Qwen with SFT-NL surpass GPT-4o's performance.

Model	Setting	FLD	FOLIO	Multi-LogiEval	ProntoQA
GPT-4o	Direct	53.0	72.4	71.0	98.8
	CoT	54.1	69.5	76.9	98.6
	Few-shot	**58.3**	**74.4**	84.4	99.0
	Logic-LM	46.9	72.1	83.3	**100**
	SymbCoT	47.6	71.6	72.1	**100**
	Sel-Inf	51.9	66.5	**84.9**	94.4
DeepSeek-R1	Direct	77.2	75.9	81.8	**100**
	CoT	77.6	78.8	79.0	**100**
	Few-shot	77.3	81.8	**84.6**	99.4
	Logic-LM	69.6	77.5	81.2	96.4
	SymbCoT	69.6	82.8	72.0	98.2
	Sel-Inf	**83.8**	**85.2**	73.1	96.0
Llama-3.1-8B-Instruct	Direct	31.7	54.7	40.5	64.6
	CoT	29.3	50.7	44.6	63.8
	Few-shot	41.0	46.5	59.4	48.9
	Logic-LM	38.3	52.5	44.4	77.6
	SymbCoT	38.1	58.8	46.3	78.8
	Sel-Inf	48.5	47.5	55.2	64.2
	LogiPT	53.3	**61.7**	57.9	76.4
	SFT-NL	**67.5**	57.1	**71.3**	99.6
	SFT-Symb-Struct	63.2	56.2	59.7	**99.8**
	SFT-Symb-Filter	66.7	54.7	50.8	91.0
	SFT-Symb-Direct	52.8	48.3	53.9	98.8
Qwen-2.5-7B-Instruct	Direct	46.6	61.1	37.0	90.6
	CoT	50.4	65.5	54.3	90.4
	Few-shot	53.2	68.5	61.3	91.1
	Logic-LM	46.6	**69.1**	27.1	85.8
	SymbCoT	22.6	57.5	**63.9**	87.0
	Sel-Inf	49.0	62.6	39.7	92.6
	LogiPT	58.6	61.7	55.6	52.4
	SFT-NL	**71.0**	62.6	64.3	**97.4**
	SFT-Symb-Struct	54.6	50.7	57.7	83.8
	SFT-Symb-Filter	54.7	55.7	61.0	96.0
	SFT-Symb-Direct	54.8	53.2	58.7	61.4

Table 2: Overall Benchmark Accuracy on four models with different settings.

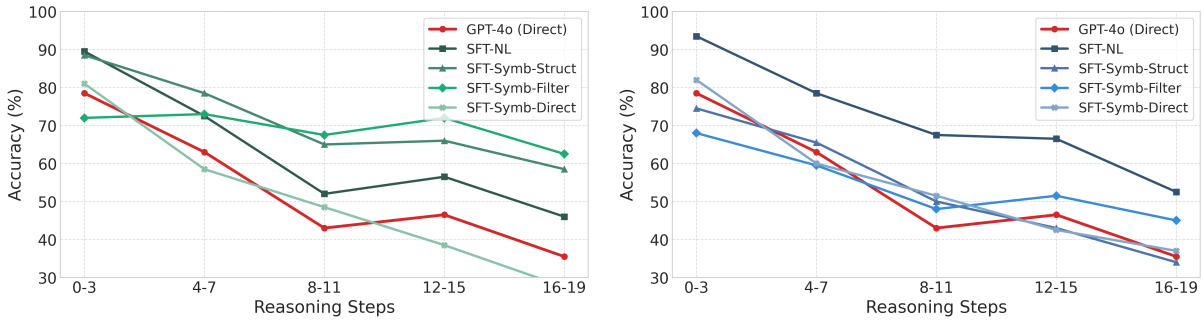
On more difficult out-of-distribution questions requiring 16–19 steps of reasoning—where no training samples are available—performance drops by approximately 10% relative to the 12–15 step range. However, even under these conditions, SFT models maintain accuracy comparable to GPT-4o. Combined with strong generalization to unseen datasets such as FOLIO and Multi-LogiEval, these results suggest that **SFT induces genuine logical reasoning ability in LLMs**. At the same time, the sharp performance decline on longer reasoning chains implies that some portion of success on shorter problems may still stem from shallow pattern matching or memorization, rather than robust inference. Detailed results can be found in C.

5.2.2 Results on Stepwise Soundness

Table 3 reports the results of **stepwise soundness evaluation** across different models and training settings, offering a more fine-grained view of how well LLMs internalize logical reasoning principles.

The **All Valid** metric measures the proportion of samples where *every* step is logically valid, a stringent indicator of formal reasoning. We observe that models fine-tuned with **SFT-NL** and **SFT-Symb-Struct** achieve high All Valid scores. Llama with SFT-NL reaches 40.9%, substantially outperforming strong baselines like GPT-4o and DeepSeek-R1. The low score of DeepSeek-R1, despite its high accuracy, highlights a key insight from FineLogic. Having been trained solely on final-answer correctness, it often compresses multiple reasoning steps and omits necessary premises for intermediate conclusions. Since guessing the final label is easier than ensuring every inference is sound, accuracy alone can be misleading. "All Valid" enforces this stricter requirement, revealing a critical gap in the model's reasoning fidelity.

The **All Relevant** metric measures the proportion



(a) Performance of Llama-3.1-8B-Instruct SFT.

(b) Performance of Qwen-2.5-7B-Instruct SFT.

Figure 3: Comparison of SFT variants' performance across different reasoning step ranges in FLD dataset. Both charts show accuracy declines with increasing inference steps, with GPT-4o (Direct) included as a reference. In (a), Llama with SFT-Symb-Filter maintains strong performance even in the 16-19 step range (out-of-distribution), while in (b), Qwen with SFT-NL shows remarkable early-stage reasoning capabilities.

Model	Setting	All Valid	All Relevant	All Atomic
GPT-4o	Few-shot	7.6	**56.2**	4.4
Deepseek-R1	Few-shot	13.1	33.8	5.7
Llama-3.1-8B-Instruct	Few-shot	4.5	17.4	1.6
	LogiPT	5.2	**28.5**	4.9
	SFT-NL	**40.9**	8.5	13.0
	SFT-Symb-Struct	35.0	15.4	24.7
	SFT-Symb-Filter	21.8	16.9	12.4
	SFT-Symb-Direct	33.7	10.2	**25.1**
Qwen-2.5-7B-Instruct	Few-shot	10.1	35.1	2.6
	LogiPT	6.4	**39.8**	5.3
	SFT-NL	27.6	5.4	8.5
	SFT-Symb-Struct	**35.3**	9.1	**19.8**
	SFT-Symb-Filter	16.7	11.7	10.5
	SFT-Symb-Direct	19.7	0.3	11.9

Table 3: Stepwise soundness of various models under settings without inference-time interventions. The best variant of Llama and Qwen is highlighted.

of samples in which every generated step is *relevant*—i.e., none of the steps are redundant or unnecessary for reaching the conclusion. GPT-4o and LogiPT perform exceptionally well on this metric, implying that they rarely generate superfluous reasoning steps. In contrast, SFT-NL and SFT-Symb-Direct consistently underperform. For SFT-NL, this may stem from the nature of natural language reasoning: due to its semantic richness and lack of structural constraints, the model may occasionally include exploratory or overly verbose steps, unsure of which inference is most effective. For SFT-Symb-Direct, the poor performance is likely due to the model may failure to fully capture interfact dependencies, resulting in reasoning sequences that are logically valid but contain unused or irrelevant steps.

The **All Atomic** metric evaluates whether every step in a reasoning chain corresponds to a single atomic inference—i.e., whether steps avoid combining multiple logical moves. Here, **SFT-Symb-Struct** consistently outperforms other settings, highlighting the advantages of structured symbolic reasoning. Symbolic reasoning is inherently more compact and constrained, which likely helps the model learn what constitutes a minimal, rule-aligned inference step. In contrast, natural language reasoning often fuses multiple reasoning rules into a single step, making it harder for the model to isolate atomic operations.

Model	Setting	CSS	RFI	NSD
Llama-3.1-8B-Instruct	–	8.0	9.9	32.0
	LogiPT	8.1	0.7	44.2
	SFT-NL	8.5	9.9	**51.5**
	SFT-Symb-Struct	8.7	11.1	36.1
	SFT-Symb-Filter	**9.7**	11.1	46.4
	SFT-Symb-Direct	9.0	**18.5**	41.2
Qwen-2.5-7B-Instruct	–	**8.6**	7.4	43.3
	LogiPT	8.1	9.2	43.2
	SFT-NL	8.2	16.0	44.3
	SFT-Symb-Struct	8.5	14.8	43.3
	SFT-Symb-Filter	8.3	16.0	**45.4**
	SFT-Symb-Direct	8.6	**18.5**	43.3

Table 4: Evaluation of Correctness Spanning Steps (CSS), Redundant Fact Identification (RFI), and Next-step Derivability (NSD) on Llama and Qwen. '-' indicates the original model. The best variant is highlighted.

5.2.3 Results on Representation-level Probing

Table 4 presents results from our probing analysis, which assesses whether models internally acquire key reasoning abilities. Our findings reveal several key takeaways:

Correctness Spanning Steps (CSS) shows a clear ceiling for 7/8B models under SFT. Across all SFT variants and base models, the CSS score fluctuates by less than 1%. This flat trend suggests that supervised fine-tuning alone offers little headroom for improving how early a model internally converges on the correct answer at this scale. Enhancing this "foresight" capability may require alternative methods like reinforcement learning with step-level rewards or architectural changes.

Targeted supervision can unlock specific internal skills, as shown by Symb-Direct on RFI. The **Redundant Fact Identification (RFI)** metric benefits most from the **SFT-Symb-Direct** setting. This training data uses minimal symbolization but retains redundant facts, forcing the model to learn to distinguish necessary from unnecessary premises. This direct alignment between the training task and the probing metric demonstrates that when a specific skill like redundancy filtering is desired, targeted supervision that mirrors the probe is highly effective.

Next-Step Derivability (NSD) remains a challenge. Performance on the **Next-Step Derivability (NSD)** task shows no consistent improvement across SFT variants. This indicates that vanilla SFT does not effectively enhance the model's internal representation of what logical steps are derivable next. This highlights a remaining gap in teaching explicit, forward-looking inference, suggesting promising future directions such as combining the NSD probe with self-supervised next-step prediction or curriculum-based RL.

6 Future Work: Guiding Faithful Reasoning with Reinforcement Learning

While supervised fine-tuning improves reasoning, reinforcement learning (RL) offers a path to refining it further. However, a major challenge in applying RL to logical reasoning is the reliance on sparse, binary rewards (i.e., whether the final answer is correct). This can lead to "reward hacking," where models learn heuristics to guess the right answer without developing a sound reasoning process.

The FineLogic framework offers a direct solution by providing dense, multi-faceted reward signals that can guide the quality of the reasoning process itself. Because our metrics are reference-free and deterministically scored, they translate cleanly into a multi-objective reward function.

A Multi-Objective Reward Framework. The FineLogic metrics can be directly integrated into a granular, multi-objective reward function for RL. The **stepwise soundness** metrics (**validity**, **relevance**, **atomicity**) can be used to directly reward logical integrity, promoting reasoning that is rigorous, concise, and transparent. Concurrently, the representation-level metric **CSS** can serve as an online reward for cognitive efficiency, encouraging the model to converge on the correct answer earlier in its internal process. These components can be combined in a weighted sum:

$$\begin{aligned} R_{\text{total}} = & R_{\text{acc}} + w_v R_{\text{valid}} + w_r R_{\text{relevant}} \\ & + w_a R_{\text{atomic}} + w_c R_{\text{css}} \end{aligned} \tag{6}$$

In this framework, practitioners can tune the weights $(w_v, w_r, \dots)$ to prioritize desired reasoning qualities like rigor or conciseness, offering a practical roadmap for training more faithful and interpretable models.

7 Conclusion

We introduce FineLogic, a unified and fine-grained framework for evaluating the logical reasoning capabilities of large language models. By integrating overall benchmark accuracy, stepwise soundness, and representation-level probing, FineLogic enables more interpretable and rigorous assessment beyond final-answer correctness. Leveraging this framework, we conduct a systematic investigation of how different fine-tuning supervision formats impact reasoning ability. Our experiments demonstrate that while natural language supervision leads to strong generalization and benchmark gains, symbolic styles better support minimal, rule-aligned reasoning structures. Furthermore, representation-level probing reveals that SFT primarily affects how models generate stepwise solutions rather than their ability to predict answers directly. These findings offer practical guidance for designing supervision strategies tailored to different reasoning objectives and highlight the importance of evaluating both behavioral and internal reasoning quality when advancing LLM reasoning systems.

Limitations

Our work has two primary limitations. First, our evaluation is conducted on a fixed suite of datasets. While these were chosen for their diversity, this necessarily scopes our findings to these specific benchmarks, and the observed model behaviors may not generalize to the entire domain of logical reasoning tasks. Second, our analysis of supervision styles treats each format in isolation. Consequently, the potential benefits of hybrid training strategies, which could combine the complementary strengths of natural language and symbolic formats, remain unexplored in this study.

Acknoledgement

We are grateful to Xiuying Chen of MBZUAI for her constructive comments and helpful suggestions, which helped improve our manuscript.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. 2023. Language models can be logical solvers. *arXiv preprint arXiv:2311.06158*.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. Glore: When, where, and how to improve llm reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*.

Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, and 1 others. 2025. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, and 1 others. 2024. Social science meets llms: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. Scemqa: A scientific college entrance level multimodal question answering benchmark. *arXiv preprint arXiv:2402.05138*.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Man Luo, Shrinidhi Kumbhar, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, Chitta Baral, and 1 others. 2023. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *arXiv preprint arXiv:2310.00836*.

Philipp Mondorf and Barbara Plank. 2024. Liar, liar, logical mire: A benchmark for suppositional reasoning in large language models. *arXiv preprint arXiv:2406.12546*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi,

and Yasuhiro Sogawa. 2023. Learning deductive reasoning from synthetic corpus based on formal logic. In *International Conference on Machine Learning*, pages 25254–25274. PMLR.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37:73572–73604.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Theo X Olausson, Alex Gu, Benjamin Lipkin, Cedegao E Zhang, Armando Solar-Lezama, Joshua B Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164*.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. *arXiv preprint arXiv:2404.15522*.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *arXiv preprint arXiv:2406.17169*.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*.

Hyun Ryu, Gyeongman Kim, Hyemin S Lee, and Eunho Yang. 2024. Divide and translate: Compositional first-order logic translation and verification for complex logical reasoning. *arXiv preprint arXiv:2410.08047*.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36:3083–3105.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2023. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. *arXiv preprint arXiv:2310.18659*.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. 2024. Verifiable, debuggable, and repairable commonsense logical reasoning via llm-based theory resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6634–6652.

Siyuan Wang, Enda Zhao, Zhongyu Wei, and Xiang Ren. 2025a. Stepwise informativeness search for improving llm reasoning. *arXiv preprint arXiv:2502.15335*.

Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. 2025b. Adareasoner: Adaptive reasoning enables more flexible thinking. *arXiv preprint arXiv:2505.17312*.

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

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*.

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

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

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. *arXiv preprint arXiv:2205.12443*.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2023. Harnessing the power of large language models for natural language to first-order logic translation. *arXiv preprint arXiv:2305.15541*.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *The Thirteenth International Conference on Learning Representations*.

Yulai Zhao, Haolin Liu, Dian Yu, SY Kung, Haitao Mi, and Dong Yu. 2025. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*.

Tong Zheng, Lichang Chen, Simeng Han, R Thomas

McCoy, and Heng Huang. 2025. Learning to reason via mixture-of-thought for logical reasoning. *arXiv preprint arXiv:2505.15817*.

Yujun Zhou, Yufei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. 2024a. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *Preprint*, arXiv:2509.15194.

Yujun Zhou, Jingdong Yang, Yue Huang, Kehan Guo, Zoe Emory, Bikram Ghosh, Amita Bedar, Sujay Shekar, Zhenwen Liang, Pin-Yu Chen, and 1 others. 2024b. Labsafety bench: Benchmarking llms on safety issues in scientific labs. *arXiv preprint arXiv:2410.14182*.

A Detailed Experimental Setup

A.1 Detailed Dataset Information

FLD (Morishita et al., 2024) This is a synthetic dataset designed to test generalization across varying reasoning depths. Our test set consists of 1,100 samples spanning 20 reasoning depths (0–19). It is specifically constructed to include out-of-distribution problems (depths 16–19) that are more difficult than those used for fine-tuning.

FOLIO (Han et al., 2022) A human-curated benchmark for first-order logic reasoning in natural language. We use the complete test set of 203 problems for our evaluation.

Multi-LogiEval (Patel et al., 2024) This dataset systematically combines multiple inference rules to create complex problems. To focus on challenging multi-step reasoning, we select 390 problems from its two most difficult subsets (depths 4 and 5), using only the First-Order and Propositional Logic categories.

ProntoQA (Saparov and He, 2022) This dataset evaluates multi-hop reasoning over a semi-synthetic knowledge base. For our test set, we follow the challenging setup from prior work (Pan et al., 2023) and use the 500 hardest 5-hop problems. (The 3-hop problems used for fine-tuning are described in Section 4).

A.2 Detailed Baseline Methods

LOGIPT (Feng et al., 2023) An approach where the language model is trained to directly generate symbolic reasoning steps, emulating a logical solver and bypassing an explicit natural language-to-symbol parsing stage.

Selection-Inference (Creswell et al., 2022) A method that performs multi-step reasoning through an iterative process, alternating between a "selection" step to identify relevant facts from the context and an "inference" step to derive new conclusions.

LogicLM (Pan et al., 2023) A neuro-symbolic framework where an LLM first translates a problem into a symbolic representation. A deterministic solver then performs the logical inference, and the LLM interprets the result back into natural language. It includes a self-refinement mechanism that uses solver feedback to correct translation errors.

SymbCoT (Xu et al., 2024) A framework that integrates symbolic logic into the Chain-of-Thought process. It translates the problem into a symbolic format and then generates reasoning steps using formal inference rules, often employing a verifier to check the logical soundness of the chain.

B Representation-Level Probing Implementation Details

We design three probing tasks to assess whether the model's internal representations capture reasoning-relevant information during multi-step logical problem solving. All probing experiments are conducted on a subset of the **FLD** dataset, specifically the 550 most complex problems requiring 10–20 reasoning steps. We use 450 problems for training and 100 for evaluation.

B.1 Representation Extraction

For all probing tasks, we extract the hidden state of the **final token from the last transformer layer** after processing the input prefix. The prefix consists of all reasoning steps up to a target step k (i.e., steps 1 to k), and the final-token representation is treated as a summary of the model's internal reasoning state at that point.

B.2 Probing Model

We use a lightweight yet effective classifier to probe the information contained in these hidden states. Specifically, we adopt a **logistic regression classifier** with feature standardization and **5-fold cross-validation** for hyperparameter selection. This setup ensures a simple and interpretable linear decision boundary while maintaining robustness against overfitting. The classifier is trained solely on the extracted representations, while the underlying language model remains frozen throughout the probing process.

B.3 Task 1: Correctness Spanning Steps

This task evaluates how early in the reasoning process the model internalizes the correct final answer. For a problem requiring n reasoning steps, we:

- Generate n input prefixes, each ending at step i, where $i \in [1, n]$.
- Train a probing classifier to predict the ground-truth label (True / False) based on the representation at each prefix.
- For each test sample, we identify the smallest i such that the classifier correctly predicts the label at step i but fails at step $i-1$.

The **correctness spanning length** is defined as $n - i$, capturing how early the model "knows" the correct answer.

B.4 Task 2: Redundant Facts Identification

This task assesses whether the model can distinguish between relevant and irrelevant facts. For each sample:

- We locate the point after all facts and the hypothesis have been presented.
- We construct six variants of the input: three with **necessary facts** (used later in the proof), and three with **redundant facts** (unused in any proof step).
- The classifier is trained to predict whether the appended facts are necessary or redundant based on the updated representation.

This task tests whether the model encodes awareness of which premises are logically relevant for solving the task.

B.5 Task 3: Next-Step Derivability

This task probes whether the model can determine which steps are logically available at a given point in the proof. For each sample:

- We randomly select six intermediate steps.
- At each step, we append three **valid next steps** (that are inferable from the current context) and three **invalid steps** (that appear later in the proof but are not yet derivable).
- The classifier is trained to distinguish between currently valid and invalid steps.

This task evaluates whether the model has encoded an implicit understanding of the forward progression of logical inference.

C Additional Experiments

This section provides further details and aggregate analysis of our experimental results.

Aggregate Findings Across Models. To broaden our findings, we evaluated two additional models, Qwen-2.5-Math-7B and Qwen-3-8B. The results, shown in Tables 5, 6 and 7, combined with our primary experiments, yield several key aggregate takeaways:

Model	Setting	FLD	FOLIO	Multi-LogiEval	ProntoQA
Qwen-2.5-7B-Math-Instruct	Few-shot	7.4	30.5	0.5	0.2
	SFT-NL	**58.8**	**56.2**	45.6	**97.0**
	SFT-Symb-Struct	38.9	49.3	**60.3**	90.6
	SFT-Symb-Filter	56.1	50.2	59.2	91.4
	SFT-Symb-Direct	47.9	43.8	39.5	51.4
Qwen-3-8B	Few-shot	44.0	**81.8**	**73.6**	**99.8**
	SFT-NL	**75.4**	64.0	48.7	99.4
	SFT-Symb-Struct	58.6	59.1	61.0	99.6
	SFT-Symb-Filter	63.6	57.1	64.1	**99.8**
	SFT-Symb-Direct	54.6	56.2	63.1	75.2

Table 5: Overall benchmark accuracy on **Qwen-2.5-7B-Math-Instruct** and **Qwen-3-8B** across training settings. Per-model column-wise maxima are in **bold**.

Model	Setting	All Valid	All Relevant	All Atomic
Qwen-2.5-7B-Math-Instruct	Few-shot	4.5	**65.8**	3.9
	SFT-NL	6.5	60.4	2.0
	SFT-Symb-Struct	20.3	21.7	11.0
	SFT-Symb-Filter	12.4	3.6	5.6
	SFT-Symb-Direct	**25.9**	0.5	**14.8**
Qwen-3-8B	Few-shot	12.1	49.5	6.5
	SFT-NL	17.3	**84.5**	3.5
	SFT-Symb-Struct	**26.1**	5.5	**13.0**
	SFT-Symb-Filter	19.1	2.3	7.6
	SFT-Symb-Direct	24.7	0.4	12.8

Table 6: Stepwise soundness after </think> on **Qwen-2.5-7B-Math-Instruct** and **Qwen-3-8B**. Per-model column-wise maxima are in **bold**.

- **Natural-language SFT is the most reliable path to higher accuracy.** Across all four base models and four datasets, SFT-NL consistently delivers the largest and most robust gains in final-answer correctness, reaffirming that aligning with the model's pre-training modality is most effective for generalization.
- **Symbolic curricula with redundant facts improve chain integrity.** SFT settings that retain logically valid but extraneous premises (Symb-Direct and Symb-Struct) tend to improve *All-Valid* and *All-Atomic* scores. This suggests that exposing models to "noisy but sound" proof environments encourages more careful and explicit rule application.
- **CSS is more sensitive to architecture and RL than to SFT.** The larger, RL-enhanced Qwen-3-8B model achieves a notably higher base CSS score than other models. However, its CSS remains largely unchanged by SFT, providing strong evidence that deeper latent reasoning depends more on model design and training regimes like RL, rather than supervised instruction tuning.

Model	Setting	CSS	RFI	NSD
Qwen-2.5-7B-Math-Instruct	–	7.4	4.9	37.1
	SFT-NL	8.6	6.2	**42.3**
	SFT-Symb-Struct	7.8	9.9	36.1
	SFT-Symb-Filter	8.3	8.6	37.1
	SFT-Symb-Direct	**8.7**	**11.6**	35.1
Qwen-3-8B	–	**10.0**	11.1	30.9
	SFT-NL	9.8	**23.5**	36.1
	SFT-Symb-Struct	9.9	9.9	32.0
	SFT-Symb-Filter	**10.0**	14.8	**39.2**
	SFT-Symb-Direct	9.7	17.3	35.1

Table 7: Representation-level probing (CSS, RFI, NSD) on **Qwen-2.5-7B-Math-Instruct** and **Qwen-3-8B**. Per-model column-wise maxima are in **bold**; the − row denotes the original model.

Detailed Analysis of Performance by Reasoning Depth. A more granular breakdown of performance by reasoning depth, presented in Table 8, reveals further nuances in these trends. While models fine-tuned with natural language supervision (e.g., Llama-3.1-SFT-NL achieving 89.5% accuracy for 0-3 steps on FLD) perform strongly on tasks with shallower reasoning depths, their symbolic reasoning counterparts tend to exhibit greater resilience as the complexity and number of reasoning steps increase. For instance, on FLD problems requiring 16-19 steps, Llama-3.1-SFT-Symb-Filter (62.5%) and Llama-3.1-SFT-Symb-Struct (58.5%) maintain higher accuracy compared to Llama-3.1-SFT-NL (46.0%), highlighting the benefit of symbolic formats for robust long-chain inference.

D Computational Resources

All supervised fine-tuning experiments were conducted using 4 NVIDIA A100 GPUs. Each model was trained for approximately 2 hours. Evalua-

Model	Setting	FLD Accuracy by Step				
		0–3	4–7	8–11	12–15	16–19
GPT-4o	Direct	78.5	63.0	43.0	46.5	35.5
	CoT	82.0	62.0	56.5	44.0	46.5
	Few-shot	81.5	68.5	53.5	46.0	47.0
	Logic-LM	68.4	52.1	31.5	28.2	22.8
	SymbCoT	69.9	52.5	32.0	26.5	24.5
	Sel-Inf	64.5	55.5	49.5	49.0	55.5
DeepSeek -R1	Direct	92.5	86.0	80.5	75.0	76.5
	CoT	92.0	86.0	78.0	77.5	73.5
	Few-shot	89.0	85.0	80.5	69.0	71.0
	Logic-LM	91.4	78.6	64.8	58.2	52.4
	SymbCoT	86.4	80.5	70.9	45.2	53.4
	Sel-Inf	93.0	88.0	84.5	79.0	75.5
Llama-3.1 -8B-Instruct	Direct	40.5	30.0	24.0	27.0	25.5
	CoT	41.5	32.0	29.0	24.0	19.5
	Few-shot	49.5	45.5	33.0	39.0	32.0
	Logic-LM	56.4	41.3	32.6	28.8	26.2
	SymbCoT	57.8	41.0	39.0	37.8	35.4
	Sel-Inf	63.5	55.5	52.5	45.0	42.0
	LogiPT	72.5	53.5	51.0	35.0	37.0
	SFT-NL	89.5	72.5	52.0	56.5	46.0
	SFT-Symb-Struct	88.5	78.5	65.0	66.0	58.5
	SFT-Symb-Filter	72.0	73.0	67.5	72.0	62.5
	SFT-Symb-Direct	81.0	58.5	48.5	38.5	27.5
Qwen-2.5 -7B-Instruct	Direct	69.0	45.5	45.0	38.5	36.0
	CoT	70.5	55.5	36.5	42.5	40.5
	Few-shot	63.0	44.0	33.5	27.0	33.0
	Logic-LM	68.7	51.2	31.4	26.0	29.2
	SymbCoT	52.3	39.5	30.7	28.1	19.9
	Sel-Inf	49.0	26.5	29.5	27.0	24.5
	LogiPT	80.5	74.0	64.0	68.0	57.5
	SFT-NL	93.5	78.5	67.5	66.5	52.5
	SFT-Symb-Struct	74.5	65.5	50.0	43.0	34.0
	SFT-Symb-Filter	68.0	59.5	48.0	51.5	45.0
	SFT-Symb-Direct	82.0	60.0	51.5	42.5	37.0

Table 8: FLD accuracy breakdown by reasoning step ranges

tion on the full suite of benchmarks and diagnostic metrics required an additional 0.5 hours per model.

E Example and Case Study

This section showcases examples from our training dataset along with an error case study. Further details can be found in Figure 4 and Figure 5.

F Prompt Template

This section showcases various prompts, encompassing those designed for reasoning and data generation, as detailed in Figures 6,7,8,9,10,11,12,13,14

G Use of AI Assistants in Manuscript Preparation

During the preparation of this manuscript, we utilized large language models, including GPT and Gemini, as writing assistants. Their role was limited to improving the clarity, grammar, and readability of the text. This included tasks such as rephrasing sentences, correcting spelling errors, and ensuring stylistic consistency. The core scientific contributions, experimental design, and analysis presented in this paper are solely the work of the human authors, who take full responsibility for the final content.

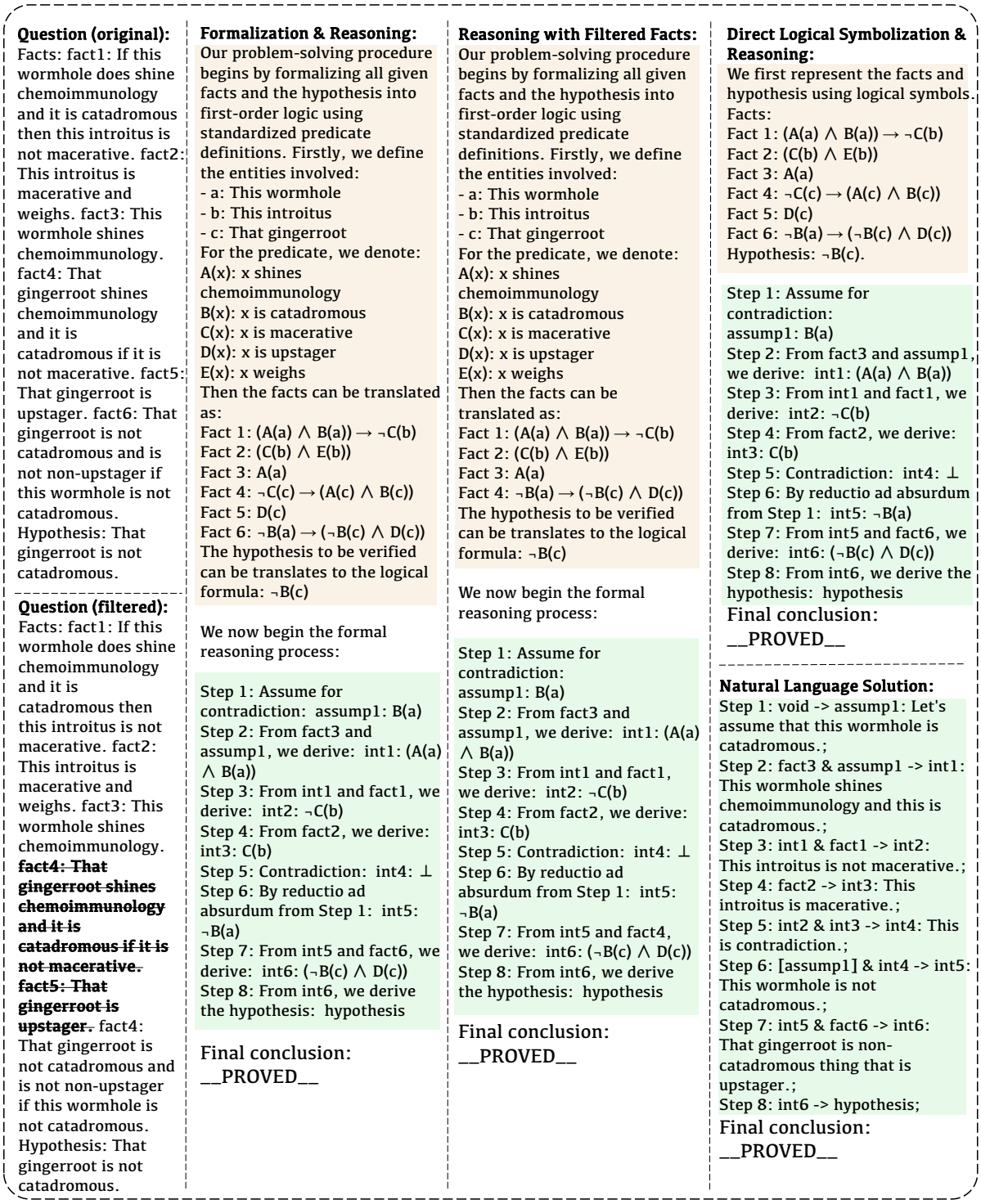


Figure 4: Comparison of a logical reasoning problem under four distinct training data settings. The figure illustrates: (a) direct logical symbolization and reasoning ; (b) full formalization in first-order logic, including definitions and fact translation ; (c) reasoning conducted purely in natural language; and (d) formal reasoning using a pre-filtered set of facts. This comparison highlights the differences in processing pathways and the structure of the resulting solutions for each approach.

Facts and Hypothesis

Fact1: Benignantness happens.
Fact2: That hands happens results in that that Highlanderness but not enlightening reincarnationism occurs is true.
Fact3: Calculating jailed takes place.
Fact4: If lavishness happens prostitution occurs but fandango does not takes place.
Fact5: That readiness but not arundinaceousness occurs are caused by forgettableness.
Fact6: If trumpeting Asvins takes place then amauroticness happens and outdoing glasnost does not occurs.
Fact7: That fandango takes place yields that lavishness does not takes place.
Fact8: Fandango brings about that prostitution but not lavishness occurs.
Fact9: Combining Yorkshire takes place.
Fact10: Fandango occurs if that closing revisal happens is correct.
Fact11: That lavishness does not occurs holds.
Fact12: Trumpeting Asvins but not stargazinging happens.
Fact13: That quarantining does not occurs leads to that that lavishness occurs and disgruntleding tularaemia does not takes place is right.
Fact14: Quarantining does not happens if that repressing Sclerodermataceae takes place and asterismalness occurs is wrong.

Hypothesis: The statement that prostitution is caused by closing revisal is incorrect."

GPT-4o Response (Non-Necessary case)

To determine if the hypothesis "The statement that prostitution is caused by closing revisal is incorrect" is correct, we must reason using the provided facts.

Step 1: Assume closing revisal happens. (assumption for contradiction)
Step 2: Fact10 states, "Fandango occurs if that closing revisal happens is correct." Thus, if closing revisal happens, then fandango takes place.
...
Step 7: Using int1 and Fact7, it follows that lavishness does not take place. (int3)
Step 8: Fact11 confirms, "That lavishness does not occurs holds," which is consistent with our finding in int3.
Step 9: We proved that fandango happening (due to closing revisal) leads to prostitution happening, supporting step-wise reasoning until prostitution indeed occurs through factual consequences.
Step 10: Therefore, the assumption that prostitution is directly caused by closing revisal through the indirect chain (through fandango) leads to a valid conclusion that prostitution occurs.
Step 11: Hence, there is no contradiction observed that refutes the original claim that closing revisal causes prostitution through the chain found in statements and facts, indirectly.
Based on this reasoning, the hypothesis stating that "prostitution is caused by closing revisal is incorrect" does not hold any demonstrable contradictions and isn't proven false with the current set of facts.

Final conclusion: __DISPROVED__

Figure 5: Case Study: In the GPT-4o response, Step 8 redundantly confirms the conclusion from Step 7 (int3) by citing Fact11, making it an unnecessary step in the reasoning chain. Step 9 is not atomic because it summarizes a multi-step inference (from "fandango happening" to "prostitution happening" using Fact 8) as a single proven outcome without detailing the individual logical operations involved.

Prompt Template: Direct Reasoning

```
Based on the provided facts, answer the question. Conclude with one
of the markers: "__PROVED__" for proven, "__DISPROVED__" for disproven, or
"__UNKNOWN__" if uncertain.
Facts:{facts}
Hypothesis:{hypothesis}
```

Figure 6: Prompt template for direct reasoning. Placeholders: `{facts}`, `{hypothesis}`.

Prompt Template: CoT Reasoning

```
Based on the provided facts, answer the question. Conclude with one
of the markers: "__PROVED__" for proven, "__DISPROVED__" for disproven, or
"__UNKNOWN__" if uncertain.
Facts:{facts}
Hypothesis:{hypothesis}
Let's analyze this step by step.
```

Figure 7: Prompt template for Chain-of-Thought (CoT) reasoning. Placeholders: `{facts}`, `{hypothesis}`.

Prompt Template: Few-Shot Reasoning

```
Based on the provided facts, answer the question. Conclude with one
of the markers: "__PROVED__" for proven, "__DISPROVED__" for disproven, or
"__UNKNOWN__" if uncertain.
Here are some examples of proofs for your reference:
[Start of example]
For example, for this question:
{example}
[End of example]
You can refer to the proof method of the above question, think step by step, and
give the result of this question.
Facts:{facts}
Hypothesis:{hypothesis}
```

Figure 8: Prompt template for few-shot reasoning. Placeholder: `{example}`, `{facts}`, `{hypothesis}`..

Prompt Template: Entity and Predicate Extraction

```
You are a logic analysis expert. Please extract all entities and predicates
from the following logical expression translations:
Translation content: {formula_translations}
facts_formula: {facts_formula}
facts: {facts}
Special Requirement: If any entity or predicate symbol appears in the
facts_formula, but has NO direct definition in the Translation content, you
MUST go to the facts section and locate the corresponding natural language
description and extract it. Be extremely careful NOT to omit any such entities
or predicates. Only skip if it is literally missing from both translation content
and facts.
Task:
1. Identify all entities involved (e.g., this tablefork, this corsair) and
assign variables to them (a, b, c, d...)
2. Identify all predicates (e.g., is a raised, is a collotype) and assign symbols
(using the original symbols like A, B, C...)
Critical instructions:
- Only give full entity and predicate explanations if their definitions appear
in the formula_translations or facts.
- Only include entities and predicates that explicitly appear in the provided
translation content or facts.
- Do not invent, infer, or add any entities or predicates not directly mentioned
in the translations or facts.
- Maintain the original variable identifiers (e.g., 'a' in A(a) corresponds to
the first entity).
- Maintain the original predicate identifiers (e.g., 'A' in A(x) represents "x
is a raised").
- If a symbol (like 'c', 'F', etc.) doesn't appear in the translations or facts,
do not include it in your output.
Expected output format:
We define the entities involved:
- a: [Corresponding entity, e.g., "This tablefork"]
- b: [Corresponding entity, e.g., "This corsair"]...
We denote:
[Original predicate symbol](x): [Predicate description]
[Original predicate symbol](x): [Predicate description]...
Please provide only the requested definitions without any additional information
or explanations.
```

Figure 9: Prompt template for extracting entities and predicates when lowercase variables (entities) are present. Placeholders: {formula_translations}, {facts_formula}, {facts}.

Prompt Template: Predicate Extraction (No Entities)

```
You are a logic analysis expert. Please extract all predicates from the
following logical expression translations:
Translation content: {formula_translations}
facts_formula: {facts_formula}
facts: {facts}
Special Requirement: If any entity or predicate symbol appears in the
facts_formula, but has NO direct definition in the Translation content, you
MUST go to the facts section and locate the corresponding natural language
description and extract it. Be extremely careful NOT to omit any such entities
or predicates. Only skip if it is literally missing from both translation content
and facts.
Task: Identify all predicates and translate each uppercase symbol directly.
Critical instructions:
- For each uppercase symbol in the facts_formula, provide a direct translation
in the format: [SYMBOL]: xxx happened.
- **Do not omit any symbols that appear in facts_formula or translation content.
If they appear, they must be translated.**
- Only include symbols that actually appear in the facts_formula or translation
content.
- Do not invent or infer any entities or relationships not explicitly mentioned.
- If a predicate's meaning is clearly defined in the translations or facts, use
that definition.
- Do not include any lowercase symbols or entity definitions as they are not
relevant in this case.
- If some symbols appear in facts_formula but not in translation content, you
can directly translate the entire formula expression containing those symbols
rather than translating each symbol individually. For example, for an expression
like ¬C→¬(F∧¬E), you don't need to separately translate E if it's not defined
elsewhere.
Expected output format:
We define:
A: xxx happened.
B: xxx happened.
AB: xxx happened...
Please provide only the requested definitions without any additional information
or explanations.
```

Figure 10: Prompt template for extracting predicates when no lowercase variables (entities) are present. Placeholders: {formula_translations}, {facts_formula}, {facts}.

Prompt Template: Logic Proof Translation

```
You are a logic proof translator. Your task is to translate a logical proof sequence from symbolic notation into a clear, step-by-step explanation.
Given: 1. A proof sequence in symbolic form 2. Definitions of entities and predicates used in the proof 3. Logical formula translations
Task: Convert the symbolic proof into a concise, step-by-step explanation that a human can easily follow.
Proof sequence to translate: {proofs_sentence}
Conclusion: {conclusion}
Instructions for translation:
1. Split the proof at each semicolon (;) to identify individual steps.
2. For each step: First, write a brief, natural language explanation on its own line (e.g. "Assume for contradiction: [formula]" or "From [inputs], we derive:"). On the next line, write the step label and the logical formula as in the original proof (e.g. assump1: A(b), int2: ¬B(b), etc.). Do not put both the explanation and the formula on the same line. For assumptions, use "Assume for contradiction: [formula]" then write assumpX: [formula] on the following line. For a standard derived step, use "From [inputs], we derive:" then on the following line write intX: [formula]. For contradictions, use "Contradiction:" then on the following line write "⊥". For reductio ad absurdum, use "By reductio ad absurdum from [step number]:" then write the derived conclusion on the next line. Do not skip formula labels or step names. Write both the explanation and the labeled formula.
3. Maintain correct logical notation (such as ¬, ∧, ∨, →, ∃, ⊥, etc.).
4. In the final step, clearly relate the conclusion to the hypothesis, if appropriate.
5. The output should be only the formatted translation, with no additional commentary.
Output format:
Step 1: [Brief explanation]
[Formula derived]
Step 2: From [input], we derive:
[Formula derived]
Step 3: Assume for contradiction:
assumpX: [Formula derived]
...
{status_message_content}
Final conclusion: {conclusion}
The conclusion must use exactly two underscores before and after either PROVED or DISPROVED or UNKNOWN, with no additional spaces or characters. Translate the proof concisely but retain all logical information from the original proof sequence. Do not add any steps not present in the original, and do not skip any steps. Output the translation only, with no additional commentary.
```

Figure 11: Prompt template for logic proof translation. The placeholder `{proofs_sentence}` is for the symbolic proof sequence. The placeholder `{conclusion}` is for the conclusion (__PROVED__/__DISPROVED__/__UNKNOWN__). The placeholder `{status_message_content}` is replaced by the string 'The search path has been exhausted without finding a way to either prove or disprove the hypothesis.' if `{conclusion}` is '__UNKNOWN__', and is an empty string otherwise (which will result in different spacing around it as per the original prompt generation logic).

Prompt Template: Logical Proof Generation

```
Solve the following logical reasoning problem using formal symbolic logic and
provide a step-by-step reasoning process.
Follow these steps precisely:
1. Define predicates to represent terms in the problem
2. Translate all facts and the hypothesis into formal logical expressions
3. Derive the conclusion through systematic reasoning
4. State the final conclusion
OUTPUT FORMAT:
Your answer should follow this format exactly:
- Begin with "Our problem-solving procedure begins by formalizing all given
facts and the hypothesis into first-order logic using standardized predicate
definitions."
- Then state "For the predicate, we denote:" followed by your predicate
definitions
- Translate each fact into a formal logical expression
- Present your reasoning steps in numbered format (Step 1:, Step 2:, etc.)
- End with "Final conclusion: " followed by either "__PROVED__" or "__DISPROVED__"
IMPORTANT: The conclusion must use exactly two underscores before and after
either PROVED or DISPROVED, with no additional spaces or characters.
Here is an example problem solution, You need to strictly follow the format like
this:
Example Solution:
{fewshot_example}
Now, solve this problem: {question}
The answer should be: {label}
Provide only the solution with no additional commentary or preamble.
```

Figure 12: Prompt template for generating a logical reasoning process. Placeholders: `{question}` for the problem statement, `{label}` for the expected answer (e.g., "__PROVED__"), and `{fewshot_example}` for a formatted example solution.

Prompt Template: Step Validity Evaluation

```
 Premises:
{premises_str}

Conclusion:
{concl_text_full}

Do the premises entail the conclusion? Answer true or false only.
```

Figure 13: Prompt template for evaluating step validity. Placeholders: `{premises_str}` (a string listing the premises, e.g., "fact1: Text of fact 1 int1: Text of intermediate 1"), `{concl_text_full}` (a string representing the conclusion, e.g., "int2: Text of intermediate 2" or "hypothesis: Text of hypothesis"). The model is expected to return 'true' or 'false'.

Prompt Template: Step Atomicity Evaluation

```
 Premises:
{premises_str}

Conclusion:
{concl_text_full}

Is this inference atomic...? Answer true or false only.
```

Figure 14: Prompt template for evaluating step atomicity. Placeholders: `{premises_str}` (a string listing the premises), `{concl_text_full}` (a string representing the conclusion). The model is expected to return 'true' or 'false' indicating if the inference from premises to conclusion is a single, indivisible logical step.