Logical Reasoning in Large Language Models: A Survey

Anonymous ACL submission

Abstract

With the emergence of advanced reasoning models like OpenAI o4 and DeepSeek-R1, large language models (LLMs) have demonstrated remarkable reasoning capabilities. However, their ability to perform rigorous logical reasoning remains an open question. This survey synthesizes recent advancements in logical reasoning within LLMs, a critical area of AI research. It outlines the scope of logical reasoning in LLMs, its theoretical foundations, and the benchmarks used to evaluate reasoning proficiency. We analyze existing capabilities across different reasoning paradigms — deductive, inductive, abductive, and analogical — and assess strategies to enhance reasoning performance, including data-centric tuning, reinforcement learning, decoding strategies, and neuro-symbolic approaches. The review concludes with future directions, emphasizing the need for further exploration to strengthen logical reasoning in AI systems.

1 Introduction

Logical reasoning is a fundamental challenge to artificial intelligence (AI) and natural language processing (NLP) (Newell and Simon, 1956; McCarthy and Hayes, 1981; McCarthy, 1959). While early formal logic-based reasoning approaches faced limitations in scalability and adaptability (Pereira, 1982; Cann, 1993), data-driven models became the dominant method since the 1980s (McCarthy, 1989). Recently, pre-trained Large Language Models (LLMs) and their emergent logical reasoning abilities have attracted increasing attention (Liu et al., 2023b; Xu et al., 2023). Logical reasoning integrates LLMs with inference structuring, enabling multistep deduction and abstraction, and improving interpretability and reliability (Shi et al., 2021; Stacey et al., 2022; Rajaraman et al., 2023). It also strengthens generalization, helping models handle novel scenarios beyond their training data (Haruta et al., 2020). As LLMs become integral to domains like legal analysis and scientific discovery, ensuring the correctness and verifiability of their reasoning is increasingly vital. As a result, post-training LLM for reasoning has garnered a surge of interest in both industry and research(OpenAI, 2024; DeepSeek-AI, 2025; Muennighoff et al., 2025).

Recent surveys have touched upon LLMs' reasoning (Li et al., 2025; Huang and Chang, 2023). However, existing surveys discuss general reasoning, exemplified by chain-of-thought (CoT), treating logical reasoning as a task case, without dedicated discussion. There has been lack of a thorough literature review focusing on LLMs and formal symbolic logic. To address this issue, this survey provides a comprehensive review of logical reasoning in large language models (LLMs), with a focus on formal and symbolic logic-based reasoning rather than general heuristic approaches. The structure is illustrated in Figure 1. We begin by defining logical reasoning in AI, distinguishing it from general-purpose reasoning, and categorizing key paradigms, including deductive, inductive, abductive, and analogical reasoning. Then, we analyze existing benchmarks and evaluation methodologies, identifying gaps in assessing symbolic inference, consistency, and robustness. We further explore state-of-the-art techniques for enhancing logical reasoning, such as supervised fine-tuning, logic-informed pre-training, reinforcement learning, inference-time decoding strategies, and hybrid neuro-symbolic methods. We examine recent advances in neuro-symbolic integration, along with applications of theorem provers, logic solvers, and formal verification frameworks in LLMs. Finally, we highlight open challenges in scalability, reasoning consistency, explainability, and efficiency, proposing future directions for multi-modal reasoning, hybrid architectures, and improved evaluation frameworks.

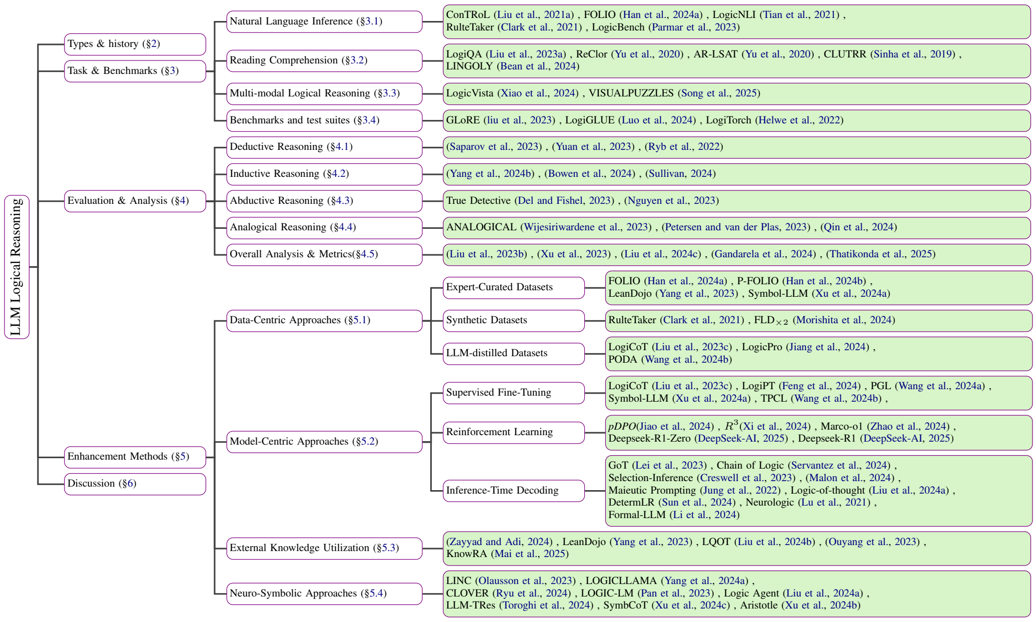

Figure 1: The structure of this survey

2 Logic in Artificial Intelligence

Logical reasoning is a cornerstone of artificial intelligence (AI), enabling machines to simulate human thought processes and solve complex problems. At its core, logical reasoning applies structured rules to derive conclusions from premises, providing a rigorous framework for decision-making and inference (Sun et al., 2023).

2.1 History of Logic Reasoning Research

Logical reasoning can be traced back to ancient Greece, where Aristotle's syllogisms laid the foundation for classical logic. During the Middle Ages, scholars refined these theories, and in the 17th century, Leibniz's universal language and calculus ratiocinator bridged logic with mathematics, foreshadowing modern computational logic. The 19th century saw George Boole's Boolean algebra, which transformed logic into a mathematical framework, laying the foundation for digital computing.

The 20th century ushered in modern logic, with Russell and Whitehead's Principia Mathematica formalizing complex logical systems. By the mid-century, AI pioneers like John McCarthy leveraged logic for knowledge representation and automated theorem proving, leading to logic programming and knowledge bases. The 1970s introduced non-monotonic logic, enabling AI to handle common-sense reasoning. The 1980s saw logical reasoning integrate with knowledge representation, advancing expert systems for real-world applications. The 1990s saw the rise of knowledge graphs, structuring vast knowledge for complex reasoning tasks.

With the development of deep learning in the 21st century, neuro-symbolic approaches stand out as a new approach for combining deep learning with logical inference, resulting in tools like DeepLogic (Cingillioglu and Russo, 2019) and SAT-Net (Wang et al., 2019). Logical reasoning remains a cornerstone of AI research, evolving from philosophy to modern computing. As AI advances, logical reasoning continues to shape intelligent systems, ensuring structured, interpretable, and robust decision-making.

2.2 Types of Logical Reasoning

Logical reasoning can be broadly categorized into four main types, each serving distinct purposes and applications:

Deductive Reasoning. This type of reasoning derives specific conclusions from general principles or premises. It operates under the rule that if all premises are true and the reasoning is valid, the conclusion must also be true. Deductive reasoning is fundamental in fields such as mathematics and formal logic, where certainty and rigor are paramount.

Inductive Reasoning. Unlike deductive reasoning, inductive reasoning draws general conclusions based on specific observations or evidence. While the conclusions are often considered probable, they are not guaranteed to be true. Inductive reasoning is widely used in scientific discovery and data-driven

Dataset	Lang.	Task	Reasoning Type	Size	Source
LOGIQA	Zh/En	MRC	Misc.	8,678	Exam
ReClor	En	MRC	Misc.	6,138	Exam
AR-LSAT	En	MRC	Misc.	2,064	Exam
CLUTRR	En	MRC	Inductive	6,016	Rule
GSM	En	MRC	Deductive	19K	Exam
LINGOLY	En	MRC	Inductive	1,133	Expert
ConTRoL	En	NLI	Deductive	8,325	Exam
FOLIO	En	NLI	Deductive	1,351	Expert
LogicNLI	En	NLI	Deductive	30K	Exam
ProofWriter	En	NLI	Deductive	-	Exam
LogicBench	En	NLI	Deductive	1,270	Rule
ART	En	NLI	Abductive	20K	Expert
Analogical	En	NLI	Analogical	720K	Crowd
GLoRE	Zh/En	Misc.	Misc.	17 tasks	Misc.
LogiGLUE	En	Misc.	Misc.	24 tasks	Misc.
LogiTorch	En	Misc.	Misc.	16 tasks	Misc.
BIG-Bench	En	Misc.	Misc.	7 tasks	Misc.

Table 1: Main Datasets and Benchmarks of Logical Reasoning Task.

decision-making, where patterns and trends are inferred from empirical data.

Abductive Reasoning. This form of reasoning seeks the most plausible explanation or cause for a set of observations, often in the presence of incomplete information. Abductive reasoning is particularly useful in diagnostic tasks and real-world problem-solving. While abductive conclusions are not certain, they provide a practical basis for hypothesis generation and decision-making under uncertainty.

Analogical Reasoning. Analogical reasoning involves drawing comparisons between similar situations or domains to make inferences or solve problems. By identifying parallels between different scenarios, this type of reasoning enables creative problem-solving and knowledge transfer. Analogical reasoning is particularly valuable in fields like education, design, and innovation.

3 Tasks and Benchmarks

Logical reasoning datasets and benchmarks are essential for evaluating the reasoning capabilities of large language models (LLMs). These datasets can be categorized into three types based on their data sources:

Rule-based Datasets (Tafjord et al., 2021; Sinha et al., 2019) are automatically generated using logical rules, enabling large-scale data collection. However, ensuring diversity is crucial to avoid repetitive patterns and comprehensively evaluate reasoning capabilities.

Crowdsourced Datasets (Bhagavatula et al., 2020) leverage collective human intelligence for diverse reasoning tasks. While requiring quality control measures, these datasets capture nuanced linguistic patterns and commonsense knowledge that automated methods often miss.

Expert-Designed Datasets (Han et al., 2024a) are constructed by domain experts, ensuring high precision and accuracy. Although typically smaller than crowd-sourced corpora, their meticulous design makes them indispensable for in-depth logical reasoning evaluation.

Exam-Based Datasets (Liu et al., 2021b; Yu et al., 2020; Wang et al., 2022) originate from standardized test questions (e.g., Chinese National Civil Service Exam, LSAT, CAT), offering high-quality, expert-crafted logic problems at scale. These datasets are widely used to evaluate reasoning in real-world scenarios.

Table 1 summarizes important datasets for logical reasoning, which typically cover tasks such as Natural Language Inference (NLI) (§3.1), Machine Reading Comprehension (MRC) (§3.2). Examples can be found in A.1.

3.1 Natural Language Inference (NLI)

NLI evaluates whether a *hypothesis* logically follows from a *premise*, directly assessing a model's reasoning ability. Labels typically fall into binary (Entailment, Non-entailment) or ternary (Entailment, Contradiction, Neutral) classifications. Some datasets use *True* and *False* labels instead. **ConTRoL** (Liu et al., 2021a), derived from recruitment exams, contains 8,325 entries labeled as *Correct*, *Incorrect*, or *Can't Say*, corresponding to entailment, contradiction, and neutral. **FOLIO** (Han et al., 2024a), an expert-constructed dataset for First-Order Logic (FOL) reasoning, consists of 1,351 entries labeled *True* or *False*. **LogicNLI** (Tian et al., 2021), with 30K entries generated via logical rules, isolates FOL-based inference from commonsense reasoning. **ProofWriter** (Tafjord et al., 2021) extends RuleTaker (Clark et al., 2021) by introducing closed-world (CWA) and open-world (OWA) assumptions, covering handcrafted domain theories and crowd-sourced paraphrased rules for linguistic and domain generalization. **LogicBench** (Parmar et al., 2023), generated by GPT-3, includes 1,270 test entries across 25 reasoning types (e.g., propositional logic, FOL) labeled *Yes* or *No*. **ART** (Bhagavatula et al., 2020) contains 20K commonsense narrative contexts and 200k explanations for abductive reasoning evaluation.

3.2 Machine Reading Comprehension (MRC)

Machine Reading Comprehension (MRC) evaluates logical reasoning by requiring models to answer questions based on a given passage, commonly formatted as multiple-choice, span extraction, or free response. **LogiQA** (Liu et al., 2023a), sourced from the Chinese Civil Service Exam, contains 15,937 Chinese and English entries targeting complex logical reasoning. **ReClor** (Yu et al., 2020), derived from the GMAT, features 6,138 English multiple-choice questions with four options. **AR-LSAT** (Wang et al., 2022), collected from the LSAT exam, includes 2,064 entries covering ordering, grouping, and allocation games with five options each. **CLUTRR** (Sinha et al., 2019) focuses on inductive reasoning for kinship relationships in short narratives, containing 6,016 entries combining entity extraction and logical inference. **LINGOLY** (Bean et al., 2024) uses Linguistic Olympiad puzzles (1,133 problems across 6 formats and 5 difficulty levels) to assess pattern identification and generalization in low-resource or extinct languages.

3.3 Multi-modal Logical Reasoning

Recently, logical reasoning combining texts and images has been explored in several research. **LogicVista** (Xiao et al., 2024) collects 448 comprehensive exam-based logical reasoning data in Visual contexts with human-annotated rationales suitable for both open-ended and multiple-choice evaluation. **VISUALPUZZLES** (Song et al., 2025) is a holistic benchmark of 1,167 exam-based puzzle-like multi-modal questions specifically designed to decouple reasoning abilities from domain knowledge.

3.4 Benchmark Suites

Benchmark suites standardize evaluation and facilitate model comparison in logical reasoning research. **GLoRE** (liu et al., 2023) provides 17 test-only datasets for few-shot and zero-shot evaluation of generalization capabilities. **LogiGLUE** (Luo et al., 2024) unifies 24 logical reasoning tasks into a sequence-to-sequence format with both training and test sets for comprehensive evaluation. **LogiTorch** (Helwe et al., 2022) offers a PyTorch-based framework with 16 datasets and model architectures for streamlined logical reasoning experiments. **BIG-bench** (Srivastava et al., 2022) includes 7 collaborative logical reasoning tasks such as Logic Grid Puzzle and Logical Fallacy Detection. LogiEval and LogiEval-Hard (Liu et al., 2025) provides a holistic testing suite with various sub-tasks for logical reasoning.

4 Evaluations

The rapid development of pre-trained language models (PLMs) necessitates rigorous evaluation of their logical reasoning capabilities. This section examines four reasoning paradigms—deductive, inductive, abductive, and analogical—while analyzing evaluation approaches and metrics.

4.1 Deductive Reasoning

Deductive reasoning, deriving specific conclusions from general premises, is crucial for automated theorem proving. Despite LLMs performing well on tasks like compositional proofs, standard benchmarks and encoding entailment relationships, they struggle with extended reasoning, hypothetical subproofs without examples, generalization, and sensitivity to syntactic variations (Saparov et al., 2023; Yuan et al., 2023; Ryb et al., 2022).

4.2 Inductive Reasoning

Inductive reasoning, which generalizes from specific instances to broader rules, is essential for tasks like hypothesis generation and pattern recognition. While Yang et al. (2024b) find that pre-trained models can serve as effective "reasoners", Bowen et al. (2024) show that even advanced LLMs struggle with simple inductive tasks in their symbolic settings. Similarly, Sullivan (2024) demonstrates that Transformer models, even after fine-tuning, fail to learn fundamental logical principles, indicating limited inductive reasoning capabilities.

4.3 Abductive Reasoning

Abductive reasoning, which seeks the most plausible explanations for observed phenomena, is crucial in fields like law and medicine. Del and Fishel (2023) highlights the challenges LLMs face in generating plausible hypotheses from incomplete information. In the legal domain, Nguyen et al. (2023) show that despite strong performance, models struggle with abductive reasoning, underscoring the complexity of this paradigm.

4.4 Analogical Reasoning

Analogical reasoning, which infers unknown information by comparing it with known information, is vital for tasks requiring creativity and knowledge transfer. Wijesiriwardene et al. (2023) introduced ANALOGICAL, a benchmark for long-text analogical reasoning. They find that as analogy complexity increases, LLMs struggle to recognize analogical pairs. Petersen and van der Plas (2023) show that models can learn analogical reasoning with minimal data, approaching human performance. However, Qin et al. (2024) question whether LLMs truly rely on analogical reasoning, discovering that random examples in prompts often achieve comparable performance to relevant examples.

4.5 Overall Analysis and Metrics

Liu et al. (2023b) evaluate GPT-4 and ChatGPT on benchmarks like LogiQA and ReClor, showing that while GPT-4 outperforms ChatGPT, both of them struggle with out-of-distribution tasks. Xu et al. (2023) introduce the NeuLR dataset and propose a framework evaluating LLMs across six dimensions: correctness, rigor, self-awareness, proactivity, guidance, and absence of hallucinations.

Metrics for Evaluating Logical Reasoning. Reasoning is fundamentally process-oriented rather than outcome-oriented (Leighton, 2003). Although traditional conclusion-based metrics like accuracy and F1 score are widely used for their simplicity and general applicability, they fall short in assessing the logical reasoning process (Mondorf and Plank, 2024). Recent studies have introduced rationale-based metrics to evaluate the reasoning trace. Structural parsing approaches (Saparov et al., 2023; Dziri et al., 2023) decompose the reasoning process into formalized representations or graphs to facilitate a more fine-grained evaluation. However, these methods are typically constrained by the requirement for specially structured reasoning texts. Other researchers have proposed interpretable quantitative metrics (Golovneva et al., 2022; Prasad et al., 2023), designing a variety of indicators to assess diverse properties of model reasoning. Nevertheless, these methods often rely on complex feature engineering and typically use metrics such as BERTScore or entropy, whose values lack clear physical interpretability. There remains a pressing need for a widely accepted and general rationale-based evaluation method.

5 Enhancement Methods

Enhancing LLMs' logical reasoning remains crucial. This section focuses on core strategies: Data-Centric Approaches (§5.1), Model-Centric Approaches (§5.2), External Knowledge Utilization (§5.3), and Neuro-Symbolic Reasoning (§5.4).

5.1 Data-Centric Approaches

Data-centric approaches enhance LLMs' reasoning capabilities by utilizing meticulously curated training datasets. Formally:

$$D^* = \arg\max_{D} R(M_D) \tag{1}$$

where:

- D: training datasets.
- M_D: model trained on D.
- R: performance evaluator.

This formulation highlights the central role of dataset optimization in data-centric approaches. In practice, data-centric methods typically involve three types of datasets: expert-curated datasets, synthetic datasets, and LLM-distilled datasets.

Expert-Curated Datasets. The FOLIO series (Han et al., 2024a,b) establish formal verification through FOL annotations, with P-FOLIO extending the complexity of reasoning chains for enhanced training. LeanDojo (Yang et al., 2023) provides 98k+ human-proven mathematical theorems. Additionally, Symbol-LLM (Xu et al., 2024a) systematically organizes 34 symbolic reasoning tasks to capture inter-symbol relationships across 20 distinct symbolic families. These datasets benefit from high-quality data, with the scale commonly limited by labor-intensive annotation.

Rule-based Synthetic Datasets. Rule-based synthetic data remains fundamental for data generation. RuleTaker (Clark et al., 2021) formalizes this through a three-phase pipeline: behavior formalization, example synthesis and linguistic equivalents generation. Similarly, Morishita et al. (2024) develops Formal Logic Deduction Diverse ($\text{FLD}_{\times 2}$), a synthetic dataset based on symbolic theory and previous empirical insights. Rule-based data generation enables large-scale, systematic data creation and fine-grained control over specific reasoning patterns. However, rule-based datasets often suffer from limited linguistic diversity, reduced realism and a gap between artificial templates and real-world inference complexity.

LLM-Distilled Datasets. Researchers employ advanced models such as GPT-4 and Deepseek-R1 for intermediate reasoning step distillation.

LogiCoT (Liu et al., 2023c) augments existing datasets with GPT4-generated reasoning chains, while LogicPro (Jiang et al., 2024) combines algorithmic problems with code solutions to create variable-guided reasoning data. To advance, Wang et al. (2024b) propose PODA, which generates contrastive analyses of correct/incorrect options through premise-oriented augmentation, enabling reasoning path differentiation via contrastive learning. These methods leverage the reasoning ability of advanced language models and tackle the lack of diversity and complexity of rule-based methods. However, training models on LLM-distilled data may lead to model collapse (Shumailov et al., 2023), causing the models to lose diversity and accuracy due to the loss of long-tail data present in real distributions.

5.2 Model-Centric Approaches

Model-Centric approaches enhance LLMs' reasoning capabilities by optimizing model parameters and decoding strategies. The formal objective is:

$$(\theta^*, S^*) = \arg\max_{\theta, S} R(M_\theta, S) \tag{2}$$

where:

- M_θ: model with learnable parameters θ.
- S: decoding strategy (e.g., chain-of-thought prompting, verification-based decoding).
- R: reasoning performance metric.

This formulation highlights the joint optimization of model parameters θ and decoding strategy S. Practical implementations can be categorized as:

- Instruction Fine-Tuning: optimizing θ.
- Reinforcement Learning: optimizing θ.
- Inference-Time Decoding: optimizing S.

Model-Centric approaches focus on directly improving models' reasoning capabilities by optimizing its internal mechanisms and decoding strategies, making them complementary to data-centric approaches.

5.2.1 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) optimizes LLMs through supervised learning on pairs of inputs and desired outputs. For example, Liu et al. (2023c) designs multi-grained logical instructions spanning diverse levels of abstraction and complexity. Similarly, Feng et al. (2024) SFT models to mimic logical solvers by replicating formal deduction reasoning processes. In addition, Xu et al. (2024a) implements two-stage symbolic fine-tuning through *Injection* (injecting symbolic knowledge) and *Infusion* (balancing symbol and NL reasoning).

To overcome SFT's over-fitting limitations, Wang et al. (2024b) enforce contrastive learning between factual/counterfactual paths with SFT. Further, Wang et al. (2024a) augments Llama models with a Program-Guided Learning Framework and logic-specific architecture adjustments.

In summary, the primary purpose of SFT in logical reasoning is generally to inject into the model the capability for specific reasoning manners such as symbolic reasoning or long CoT reasoning. Yue et al. (2025) demonstrates that SFT is able to introduce new reasoning patterns that are not present in the base model. However, high-quality SFT data relies on distillation from more advanced models or human annotation, which is more general and effective for expanding smaller models' boundaries. For advanced large models, obtaining and scaling up higher-quality SFT data is often challenging.

5.2.2 Reinforcement Learning

Reinforcement learning (RL) has become pivotal in optimizing large language models (LLMs), particularly since the breakthrough of Reinforcement Learning from Human Feedback (RLHF). Jiao et al. (2024) leverage RL for planning-based reasoning optimization, while Xi et al. (2024) develop R^3, achieving process supervision benefits through outcome-only supervision.

The success of large-scale RL in OpenAI-o1 (OpenAI, 2024) has inspired numerous studies. RL algorithms train o1-style models to enhance Chain-of-Thought (CoT) reasoning, addressing issues like formulaic outputs and limited long-form reasoning. For instance, Zhao et al. (2024) integrates CoT instruction fine-tuning with Monte Carlo Tree Search (MCTS) decoding for multi-path reasoning exploration. In contrast, Zhang et al. (2024) employs MCTS to generate code-reasoning data for supervised fine-tuning (SFT) and Direct Preference Optimization (DPO).

A significant breakthrough comes from DeepSeek-R1 (DeepSeek-AI, 2025), which pioneers a novel RL strategy to enhance logical reasoning. DeepSeek-R1-Zero, trained purely through RL without SFT, demonstrates impressive reasoning capabilities but faces challenges in readability and language consistency. To address this, DeepSeek-R1 introduces minimal long-CoT SFT data as a cold start before RL, achieving a balance between usability and reasoning performance. By iteratively synthesizing high-quality reasoning data through RL, DeepSeek-R1 overcomes limitations imposed by human annotators, addressing issues such as mechanistic responses, repetitive patterns, and insufficient long-chain reasoning. This approach represents a potential paradigm shift in logical reasoning optimization, pushing the boundaries of what LLMs can achieve in structured reasoning tasks. However, RL enhances sampling efficiency of existing reasoning paths encoded in the base model but does not generate new reasoning patterns (Yue et al., 2025) and therefore relies on the capabilities of the foundation models. From this perspective, SFT with high-quality human-annotated or LLM-distilled data is a simpler and more effective way to expand the boundaries of smaller models than RL.

5.2.3 Inference-Time Decoding

We categorize logical reasoning enhancement methods during inference-time into inference-time scaling and constrained decoding.

Inference-time scaling employs computational augmentation without parameter updates. One common approach is decoding with structured outputs and modular workflows. GoT (Lei et al., 2023) creates structured reasoning nodes to improve complex multi-step logical reasoning. Similarly, Chain of Logic (Servantez et al., 2024) introduces a method that first divides the legal question into smaller sub-questions for solving, and then assembles the answers to resolve the original problem. In other contexts, researchers design more complex modular workflows for better performance (Creswell et al., 2023; Malon et al., 2024).

Another inference-time scaling approach involves stimulating autonomous reasoning, guiding LLMs to iteratively refine their answers. Maieutic Prompting (Jung et al., 2022) eliminates contradictions through recursive reasoning. Similarly, Logic-of-Thoughts (Liu et al., 2024a) and DetermLR (Sun et al., 2024) progressively approach the answers in an iterative style.

Inference-time scaling offers the flexibility to improve model performance without additional training, but usually incurs higher inference costs.

Constrained decoding methods, on the other hand, focus on improving the controllability and reliability of reasoning processes. Neurologic (Lu et al., 2021) enforces predicate logic constraints, while Formal-LLM (Li et al., 2024) integrates automata for constraining plan generation.

5.3 External Knowledge Utilization

LLMs often generate incorrect answers due to hallucinations when performing complex tasks such as logical reasoning, making it necessary to incorporate external knowledge to assist in producing accurate responses. Formally, the optimal integration of external knowledge can be formulated as a joint optimization problem:

$$(M^*, K^*) = \arg\max_{M,K} R(M, K) \tag{3}$$

where:

- M: the neural model, which includes both the model's parameters and its decoding strategies (generally with the parameters unchanged).
- K: knowledge integration strategy, including knowledge source curation, structured knowledge representation, retrieval-augmented mechanisms, etc.
- R: reasoning performance evaluator.

Zayyad and Adi (2024) and Yang et al. (2023) extract data from Lean, a mathematical proof tool, to aid theorem proving. In contrast, "Logic-Query-of-Thoughts" (LQOT) (Liu et al., 2024b) decomposes complex logical problems into easier sub-questions before integrating knowledge graphs.

In reading comprehension, Ouyang et al. (2023) construct supergraphs to address complex contextual reasoning, while KnowRA (Mai et al., 2025) autonomously determines whether to accept external knowledge to assist document-level relation extraction.

5.4 Neuro-Symbolic Approaches

Neural-symbolic hybrid methods represent a burgeoning research area that aims to combine the powerful representational capabilities of deep learning with the precision and interpretability of symbolic reasoning.

Formally, a neural-symbolic hybrid system aims to optimize both the neural model M and the symbolic solver P (where P represents the symbolic reasoning process) to maximize logical reasoning performance. The overall objective can be expressed as:

$$(M^*, P^*) = \arg\max_{M,P} R(P(M(x))),$$

where:

- M: The neural model, which includes both the model's parameters and its decoding strategies. It maps the input x (e.g., natural language) into a symbolic representation z within a formal language $\mathcal{L}$:

$$z = M(x), \quad z \in \mathcal{L}.$$

- P: The symbolic solver, which operates on the symbolic representation z produced by M to generate the final output y:

$$y = P(z).$$

- R: The reasoning performance metric.

Two key directions of the optimization process:

- Improving M: including refining the model's parameters and decoding strategies to produce symbolic representations that are both accurate and compatible with P.
- Enhancing P: involving improving the symbolic solver's ability to process.

By jointly optimizing M and P, neural-symbolic hybrid systems aim to leverage the strengths of both neural networks and symbolic reasoning to achieve superior logical reasoning capabilities. It is worth noting that in earlier neural-symbolic pipelines, P is often implemented as a fixed external logical reasoning engine, and thus is generally not optimized. However, in advanced practice, LLMs are increasingly being used to perform the role of P, enabling diverse optimization.

Fundamentally, these methods involve translating problems into symbolic representations with LLMs, and external symbolic solvers solving them. For example, in LINC (Olausson et al., 2023), LLMs convert natural language (NL) into first-order logic (FOL) expressions, and utilize an external theorem prover for deductive inference.

Further efforts focus on improving NL-to-symbolic translation. One prevailing approach is directly optimizing translation through training (Yang et al., 2024a) or decoding strategies (Ryu et al., 2024), while the others depend on verification or correction mechanisms (Yang et al., 2024a; Pan et al., 2023).

Building upon these, recent advancements address the traditional pipeline limitations by fully integrating LLMs into reasoning processes. Logic Agent (LA) (Liu et al., 2024a) replaces external solvers with rule-guided LLM inference chains, while LLM-TRes (Toroghi et al., 2024) implements self-contained verifiable reasoning without external symbolic solvers. SymbCoT (Xu et al., 2024c) coordinates translation, planning, solving and verification entirely through LLMs. Xu et al. (2024b) propose Aristotle, which further systematizes the symbolic reasoning pipeline through three LLM-driven components: Logical Decomposer, Logical Search Router, and Logical Resolver. However, these modular approaches increase system complexity and do not fundamentally enhance the models' intrinsic reasoning capabilities.

6 Discussion

The landscape of logical reasoning in LLMs presents several unresolved challenges that merit deeper examination. In Appendix A.2, we provide an extensive discussion on fundamental tensions shaping current research: the gap between surface-level pattern matching and genuine logical competence; inconsistencies in robustness across varied reasoning contexts; competing priorities of interpretability versus performance in hybrid approaches; limitations in current evaluation paradigms; future directions.

7 Conclusion

This survey synthesizes the rapid advancements and persistent challenges in logical reasoning for large language models (LLMs). While LLMs demonstrate impressive heuristic reasoning, rigorous logical inference remains inconsistent due to limitations in robustness, generalization, and interpretability. We analyzed strategies to enhance reasoning, including neuro-symbolic integration, data-centric tuning, reinforcement learning, test-time scaling and other improved decoding methods, and highlighted benchmarks like FOLIO and LogiQA for systematic evaluation. Future progress hinges on hybrid architectures that unify neural and symbolic reasoning, robust evaluation frameworks, scalable methods for cross-domain and multimodal inference, and directly enhancing base model causality. Addressing these challenges will advance LLMs toward reliable, interpretable reasoning critical for real-world applications.

Limitations

We focus on formal logical reasoning, which is in line with the symbolic approaches actively studied in the literature. For general reasoning or other specific reasoning types, readers may refer to complementary surveys.

References

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. 2025. How likely do LLMs with CoT mimic human reasoning? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7831–7850, Abu Dhabi, UAE. Association for Computational Linguistics.

Andrew M Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, and 1 others. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. *arXiv preprint arXiv:2406.06196*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In *Proc. of ACL Findings*, pages 323–339.

Ronnie Cann. 1993. *Formal semantics: an introduction*. Cambridge University Press, United States.

Nuri Cingillioglu and Alessandra Russo. 2019. Deeplogic: Towards end-to-end differentiable logical reasoning. *Preprint*, arXiv:1805.07433.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2021. Transformers as soft reasoners over language. In *Proc. of IJCAI*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *Proc. of ICLR*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Technical report.

Maksym Del and Mark Fishel. 2023. True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 314–322.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, and 1 others. 2024. Language models can be deductive solvers. In *Proc. of ACL Findings*, pages 4026–4042.

João Pedro Gandarela, Danilo S Carvalho, and André Freitas. 2024. Inductive learning of logical theories with llms: A complexity-graded analysis. *arXiv preprint arXiv:2408.16779*.

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.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, and 1 others. 2024a. FOLIO: Natural language reasoning with first-order logic. In *Proc. of EMNLP*, pages 22017–22031.

Simeng Han, Aaron Yu, Rui Shen, Zhenting Qi, Martin Riddell, and 1 others. 2024b. P-FOLIO: Evaluating and improving logical reasoning with abundant human-written reasoning chains. In *Proc. of EMNLP Findings*, pages 16553–16565.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. In *Proc. of ACL*, pages 263–270.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *Proc. of EMNLP*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jin Jiang, Yuchen Yan, Yang Liu, Yonggang Jin, Shuai Peng, and 1 others. 2024. Logicpro: Improving complex logical reasoning via program-guided learning. *arXiv preprint arXiv:2409.12929*.

Fangkai Jiao, Chengwei Qin, Zhengyuan Liu, Nancy F. Chen, and Shafiq Joty. 2024. Learning planning-based reasoning by trajectories collection and process reward synthesizing. In *Proc. of EMNLP*, pages 334–350.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, and 1 others. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proc. of EMNLP*, pages 1266–1279.

Bin Lei, Chunhua Liao, Caiwen Ding, and 1 others. 2023. Boosting logical reasoning in large language models through a new framework: The graph of thought. *arXiv preprint arXiv:2308.08614*.

Jacqueline P. Leighton. 2003. *Defining and Describing Reason*. Cambridge University Press, Cambridge, UK.

Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. 2024. Formal-llm: Integrating formal language and natural language for controllable llm-based agents. *arXiv preprint arXiv:2402.00798*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021a. Natural language inference in context - investigating contextual reasoning over long texts. *Proc. of AAAI*, pages 13388–13396.

Hanmeng Liu, Yiran Ding, Zhizhang Fu, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. 2025. Evaluating the logical reasoning abilities of large reasoning models. *Preprint*, arXiv:2505.11854.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, and 1 others. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 2947–2962.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Evaluating the logical reasoning ability of chatgpt and gpt-4. *Preprint*, arXiv:2304.03439.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023c. Logicot: Logical chain-of-thought instruction tuning. In *Proc. of EMNLP Findings*, pages 2908–2921.

Hanmeng liu, Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Glore: Evaluating logical reasoning of large language models. *Preprint*, arXiv:2310.09107.

Hanmeng Liu, Zhiyang Teng, Chaoli Zhang, and Yue Zhang. 2024a. Logic agent: Enhancing validity with logic rule invocation. *Preprint*, arXiv:2404.18130.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021b. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning.

Lihui Liu, Zihao Wang, Ruizhong Qiu, Yikun Ban, Eunice Chan, and 1 others. 2024b. Logic query of thoughts: Guiding large language models to answer complex logic queries with knowledge graphs. *Preprint*, arXiv:2404.04264.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2024c. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proc. of NAACL*, pages 4288–4299.

Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, and 1 others. 2024. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *Preprint*, arXiv:2310.00836.

Chengcheng Mai, Yuxiang Wang, Ziyu Gong, Hanxiang Wang, and Yihua Huang. 2025. Knowra: Knowledge retrieval augmented method for document-level relation extraction with comprehensive reasoning abilities. *Preprint*, arXiv:2501.00571.

Christopher Malon, Martin Min, Xiaodan Zhu, and 1 others. 2024. Exploring the role of reasoning structures for constructing proofs in multi-step natural language reasoning with large language models. In *Proc. of EMNLP*, pages 15299–15312.

J. McCarthy and P.J. Hayes. 1981. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in Artificial Intelligence*, pages 431–450.

John McCarthy. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*.

John McCarthy. 1989. Artificial intelligence, logic and formalizing common sense. *Philosophical Logic and Artificial Intelligence*, pages 161–190.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In *Proc. of NeurIPS*, pages 73572–73604.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, and 1 others. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

A. Newell and H. Simon. 1956. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*.

Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023. How well do sota legal reasoning models support abductive reasoning? *Preprint*, arXiv:2304.06912.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, and 1 others. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proc. of EMNLP*, pages 5153–5176.

OpenAI. 2024. Learning to reason with LLMs. Technical report.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2023. Fact-driven logical reasoning for machine reading comprehension. *Preprint*, arXiv:2105.10334.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Proc. of EMNLP Findings*, pages 3806–3824.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, and 1 others. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proc. of ACL*, pages 13679–13707.

Mihir Parmar, Neeraj Varshney, Nisarg Patel, Santosh Mashetty, Man Luo, and 1 others. 2023. Logicbench: A benchmark for evaluation of logical reasoning.

Fernando Carlos Neves Pereira. 1982. Logic for natural language analysis.

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. In *Proc. of EMNLP*, pages 16414–16425.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.

Chengwei Qin, Wenhan Xia, Tan Wang, Fangkai Jiao, Yuchen Hu, and 1 others. 2024. Relevant or random: Can llms truly perform analogical reasoning? *Preprint*, arXiv:2404.12728.

Kanagasabai Rajaraman, Saravanan Rajamanickam, and Wei Shi. 2023. Investigating transformer-guided chaining for interpretable natural logic reasoning. In *Proc. of ACL Findings*, pages 9240–9253.

Samuel Ryb, Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. AnaLog: Testing analytical and deductive logic learnability in language models. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 55–68.

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, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, and 1 others. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. In *Proc. of NeurIPS*, pages 3083–3105.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. In *Proc. of ACL Findings*, pages 2721–2733.

Jihao Shi, Xiao Ding, Li Du, Ting Liu, and Bing Qin. 2021. Neural natural logic inference for interpretable question answering. In *Proc. of EMNLP*, pages 3673–3684.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. *Empirical Methods of Natural Language Processing (EMNLP)*.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *Preprint*, arXiv:2504.10342.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust NLI models. In *Proc. of EMNLP*, pages 3809–3823.

Michael Sullivan. 2024. It is not true that transformers are inductive learners: Probing NLI models with external negation. In *Proc. of EACL*, pages 1924–1945.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, and 1 others. 2024. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. In *Proc. of ACL*, pages 9828–9862.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, and 1 others. 2023. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Proc. of ACL Findings*, pages 3621–3634.

Ramya Keerthy Thatikonda, Wray Buntine, and Ehsan Shareghi. 2025. Assessing the alignment of fol closeness metrics with human judgement. *arXiv preprint arXiv:2501.08613*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proc. of EMNLP*, pages 3738–3747.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. 2024. Verifiable, debuggable, and repairable commonsense logical reasoning via llm-based theory resolution. In *Proc. of EMNLP*, pages 6634–6652.

Chen Wang, Xudong Li, Haoran Liu, Xinyue Wu, and Wanting He. 2024a. Efficient logical reasoning in large language models through program-guided learning. *Authorea Preprints*.

Chenxu Wang, Ping Jian, and Zhen Yang. 2024b. Thought-path contrastive learning via premise-oriented data augmentation for logical reading comprehension. *arXiv preprint arXiv:2409.14495*.

Po-Wei Wang, Priya L. Donti, Bryan Wilder, and Zico Kolter. 2019. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *Preprint*, arXiv:1905.12149.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, and 1 others. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, and 1 others. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Proc. of ACL Findings*, pages 3534–3549.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, and 1 others. 2024. Training large language models for reasoning through reverse curriculum reinforcement learning. In *Proc. of ICML*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *Preprint*, arXiv:2407.04973.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *Preprint*, arXiv:2306.09841.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, and 1 others. 2024a. Symbol-LLM: Towards foundational symbol-centric interface for large language models. In *Proc. of ACL*, pages 13091–13116.

Jundong Xu, Hao Fei, Meng Luo, Qian Liu, Liangming Pan, and 1 others. 2024b. Aristotle: Mastering logical reasoning with a logic-complete decompose-search-resolve framework. *arXiv preprint arXiv:2412.16953*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024c. Faithful logical reasoning via symbolic chain-of-thought. In *Proc. of ACL*, pages 13326–13365.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, and 1 others. 2023. Leandojo: theorem proving with retrieval-augmented language models. In *Proc. of ICONIP*.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024a. Harnessing the power of large language models for natural language to first-order logic translation. In *Proc. of ACL*, pages 6942–6959.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, and 1 others. 2024b. Language models as inductive reasoners. In *Proc. of EACL*, pages 209–225.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *Proc. of ICLR*.

Zhangdie Yuan, Songbo Hu, Ivan Vulić, Anna Korhonen, and Zaiqiao Meng. 2023. Can pretrained language models (yet) reason deductively? In *Proc. of EACL*, pages 1447–1462.

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

Majd Zayyad and Yossi Adi. 2024. Formal language knowledge corpus for retrieval augmented generation. *Preprint*, arXiv:2412.16689.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023(8).

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, and 1 others. 2024. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, and 1 others. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

A Appendix

A.1 Data Examples

Figure 2 presents example questions from datasets related to logical reasoning, serving as a supplement to the description of task types in the main text.

A.2 Further Discussion

The integration of logical reasoning into large language models (LLMs) remains a critical challenge, marked by persistent gaps between heuristic performance and formal logical rigor. Below, we analyze three unresolved tensions dominating the field and outline future directions.

Superficial Reasoning vs. Genuine Logical Competence. Despite promising results on benchmark datasets, an fundamental question persists: do LLMs truly possess logical reasoning abilities, or do they merely approximate reasoning through sophisticated pattern recognition? Some recent studies have observed that current LLMs often lack substantive causal reasoning (Bao et al., 2025; Zečević et al., 2023), which in turn limits both their capacity for genuine inference and their ability to generalize, posing a crucial challenge to be tackled.

Robustness vs. Generalization. LLMs exhibit inconsistent performance in structured reasoning tasks such as deductive inference and abductive hypothesis generation. While models fine-tuned on datasets like FOLIO (Han et al., 2024a) excel in controlled settings, they struggle with adversarial perturbations or semantically equivalent rephrasings. This inconsistency arises from their reliance on surface-level statistical correlations rather than causal relationships, coupled with limited out-of-distribution generalization. A key question persists: can LLMs achieve human-like robustness without sacrificing cross-domain adaptability? Current methods prioritize narrow task performance, leaving real-world applicability uncertain.

Interpretability vs. Performance. A central tension lies in balancing neural scalability with symbolic precision. Neuro-symbolic approaches like Logic-LM (Pan et al., 2023) and Symbol-LLM (Xu et al., 2024a) embed formal logic solvers into neural architectures, improving interpretability through step-by-step proofs. However, these methods face scalability bottlenecks with large knowledge bases or complex rule dependencies. Conversely, data-driven methods (e.g., instruction tuning on LogicBench (Parmar et al., 2024)) achieve broader task coverage but fail to generalize beyond syntactic patterns. How can we reconcile transparent reasoning with black-box model performance? Hybrid architectures offer promise but introduce computational overhead, limiting practical deployment.

> **Passage:** For a television program about astrology, investigators went into the street and found twenty volunteers born under the sign of Gemini who were willing to be interviewed on the program and to take a personality test. The test confirmed the investigators' personal impressions that each of the volunteers was more sociable and extroverted than people are on average. This modest investigation thus supports the claim that one's astrological birth sign influences one's personality.
> **Question:** Which one of the following, if true, indicates the most serious flaw in the method used by the investigators?
> A. People born under astrological signs other than Gemini have been judged by astrologers to be much less sociable than those born under Gemini.
> B. There is not likely to be a greater proportion of people born under the sign of Gemini on the street than in the population as a whole.
> C. People who are not sociable and extroverted are not likely to agree to participate in such an investigation.
> D. The personal impressions the investigators first formed of other people have tended to be confirmed by the investigators' later experience of those people.

(a) A multi-choice reading comprehension example from the LogiQA dataset.

> **Premise:** Ten new television shows appeared during the month of September. Five of the shows were sitcoms, three were hourlong dramas, and two were news-magazine shows. By January, only seven of these new shows were still on the air. Five of the shows that remained were sitcoms.
> **Hypothesis:** At least one of the shows that were cancelled was an hourlong drama.
> **Label:** Entailment

(b) An NLI example from the ConTRoL dataset.

Figure 2: Example tests of Logical reasoning in NLP tasks.

Evaluation Rigor. Existing benchmarks like LogiQA (Liu et al., 2021b) and ReClor (Yu et al., 2020) conflate reasoning ability with pattern recognition through multiple-choice formats. While efforts like NeuLR (Xu et al., 2023) curate “neutral" content to isolate reasoning from domain knowledge, they lack scope for holistic evaluation. Current metrics (e.g., accuracy, BLEU) fail to assess consistency (invariance to logically equivalent inputs) or soundness (adherence to formal proof structures). What defines a gold standard for logical reasoning evaluation? Benchmarks must prioritize systematic testing of core principles (e.g., transitivity, contraposition) over task-specific performance.

Future Directions. Addressing these challenges requires hybrid architectures that dynamically integrate neural and symbolic components, such as differentiable theorem provers, to balance scalability and precision. Equally important is the development of evaluation frameworks that stress-test models on perturbed logical statements (e.g., negated premises, swapped quantifiers) to isolate reasoning from memorization. Multi-modal reasoning, which grounds inference in diverse modalities (text, images, code), presents untapped potential for enhancing robustness and interpretability. Finally, interdisciplinary collaboration—leveraging insights from formal logic, cognitive science, and machine learning—will be essential to design systems that reason with and about uncertainty. Until LLMs reliably disentangle logic from lexicon, their deployment in high-stakes domains will remain precarious. Bridging this gap demands rigorous benchmarks, scalable hybrid methods, and a redefinition of evaluation paradigms.