Formal Proof Generation in Deductive Logic Systems with Open LLMs: An Empirical Study on Natural Deduction and Analytical Tableau

Luana Stanz Oliveira Magalhães[1], **Taís Rodrigues Sandes**[1]
Victor Aguiar Evangelista de Farias[2], **Maria Viviane de Menezes**[1]
Carlos de Oliveira Caminha Neto[2], **Davi Romero de Vasconcelos**[2]

[1]Universidade Federal do Ceará - Campus Quixadá, CE – Brazil

[2]Universidade Federal do Ceará - Campus Fortaleza, CE – Brazil

{luanastanz,tais.sandes}@alu.ufc.br

{victor.farias,vivianemenezes, caminha, davi}@ufc.br

Abstract. *Large Language Models (LLMs) have expanded the possibilities of automated reasoning, motivating interest in their ability to perform structured inferences. Logical deductive systems such as Natural Deduction and Analytical Tableau impose strict syntactic and semantic constraints, making them suitable benchmarks for evaluating formal reasoning capabilities. Despite recent progress, LLMs continue to exhibit notable limitations in constructing complete and formally valid proofs. In this paper, we investigate the performance of 7 small and medium-sized open-source LLMs (Phi-4, gpt-oss-20B, Mistral-small-3.2-24B-instruct-2506, Gemma-4-26B, Qwen3-32B, Qwen3.5-35B and gpt-oss-120B) in generating logical proofs within these two deductive systems. Proofs are generated from syntax based and rule extended prompts, in both zero shot and few shot settings. The correctness of the generated proofs is then assessed through formal verification using external proof checkers grounded in the inference rules of Natural Deduction and Analytical Tableau, ensuring rigorous rule-based validation. Our findings suggest that, in their current state, while models are capable of generating formally valid proofs under certain conditions, overall performance remains limited. Few-shot prompting consistently improves proof generation accuracy, and prompts augmented with explicit rule descriptions generally outperform syntax-only formulations. Furthermore, model size alone is not a reliable predictor of success, as smaller models occasionally approach or surpass the performance of larger alternatives.*

1. Introduction

Large Language Models (LLMs) exhibit persistent limitations in logical reasoning, particularly in tasks that require compositional generalization, global consistency and strict adherence to formal inference rules [Lin et al. 2025, Sprague et al. 2024, Wang et al. 2024]. Even when models produce step by step explanations, these reasoning chains are often only superficially plausible and may contain subtle logical inconsistencies, reflecting statistical pattern matching rather than principled deductive inference [Cobbe et al. 2021,

Creswell et al. 2023]. The work by [Arkoudas and Batzoglou 2025] highlights that models may arrive at correct answers for the wrong reasons, exploiting superficial patterns instead of performing genuine reasoning. This observation motivates evaluations that go beyond answer accuracy and assess whether the reasoning process itself satisfies formal deductive constraints.

Moreover, most evaluations of logical reasoning in LLMs rely on datasets formulated in natural language, such as FOLIO [et al. 2024a], ProofWriter [Tafjord et al. 2021], and P-FOLIO [et al. 2024b]. In these settings, semantic ambiguity and the use of world knowledge make it difficult to distinguish genuine logical reasoning from heuristic inference. For instance, the sentence "*if it rained then the ground is wet*" can be formalized by defining P: "*it rained*" and Q: "*the ground is wet*", resulting in the logical formula $P \rightarrow Q$. While the symbolic form explicitly fixes the logical structure, the natural language version allows multiple interpretations and can be solved using world knowledge without explicitly applying formal inference rules. Consequently, it is hard to determine whether the model is truly performing logical inference or exploiting linguistic regularities. Additionally, evaluation is typically limited to the correctness of the final answer, without verifying the validity of intermediate steps. In this context, the generation of formal proofs constitutes a rigorous benchmark for evaluating logical reasoning. Deductive systems provide formal frameworks for constructing proofs that establish whether conclusions follow logically from a set of premises. These systems require precise control of scope, correct application of inference rules and well structured derivations.

Recently, [Arkoudas and Batzoglou 2025] identified *formal reasoning* as a distinct category of reasoning benchmarks characterized by symbolic representations and verifiable proof construction. Because the underlying logic is fully specified, these benchmarks enable unambiguous assessment of correctness and provide a controlled setting for evaluating deductive reasoning through formal verification. Within this framework, the authors introduced the benchmark ProofGrid. In their evaluation, the authors consider a wide range of state of the art LLMs, including proprietary large scale systems such as GPT-5, GPT-4.1, GPT-4o, o3, o4-mini, Gemini-2.5-Pro, Grok-3/4, Claude Sonnet-4 and Opus-4, as well as open models such as LLaMA and DeepSeek. However, ProofGrid focuses exclusively on Natural Deduction for proof construction, and its treatment of First Order Logic is limited to proof verification tasks rather than proof generation.

In this work, we evaluate complete proofs using external proof checkers grounded in Natural Deduction and Analytical Tableau. While prior studies [McGinness and Baumgartner 2024] have employed automated theorem provers and formal verification tools enhance LLM reasoning, our focus is on investigating the capabilities of small and medium sized LLMs to generate complete and formally valid proofs. Moreover, many prior evaluations focus on final answers or isolated intermediate reasoning steps [Tafjord et al. 2021, Saparov and He 2023, et al. 2024b], which may overlook errors in the global structure of an argument and overestimate reasoning capabilities by rewarding locally plausible inferences.

In contrast, our evaluation framework requires models to construct derivations whose validity is verified automatically. This more closely reflects reasoning tasks in mathematics, science, and engineering, where the quality of a solution depends not only on reaching the correct conclusion but also on providing a coherent and logically valid

justification. Although proof checking does not provide fine-grained error classification, it enables a rigorous, objective, and fully reproducible assessment of deductive reasoning.

Furthermore, recent formal reasoning benchmarks [Arkoudas and Batzoglou 2025] have focused on frontier proprietary systems. In recent years, there has been growing interest in open source small to medium sized LLMs [Abdin et al. 2024, Mistral AI 2025, Abd and et al. 2026], not merely as lower cost alternatives, but as tools for reproducibility and controlled experimentation. These models enable systematic analysis of reasoning behavior under well defined conditions. However, their capabilities in tasks requiring logical reasoning under formal deductive constraints remain underexplored. Some approaches propose transferring reasoning capabilities through supervised learning, in which LLMs are used to generate demonstrations that serve as training data for smaller models [Shridhar et al. 2023, Wang et al. 2023, Mitra et al. 2023, Lu et al. 2024, Li et al. 2024a, Li et al. 2024b]. However, these approaches do not guarantee the logical validity of the generated demonstrations. Since LLMs are also prone to hallucinations, errors in reasoning chains may be incorporated and amplified during training. Moreover, recent evidence suggests that increasing model scale does not resolve these limitations: performance in logical reasoning tasks degrades with problem complexity, and even large models struggle to maintain global consistency in structured inference settings [Lin et al. 2025].

We make the following contributions in this paper: (i) evaluation of proof generation in both Natural Deduction and Analytical Tableau,two complementary proof paradigms that emphasize constructive derivation and systematic branch expansion, respectively, thereby providing a broader assessment of formal reasoning capabilities; (ii) investigation on proof construction in both Propositional Logic and First-Order Logic, extending evaluation beyond the propositional setting; (iii) analysis of prompting strategies on proof generation by comparison of syntax only and syntax plus rules prompts under both zero-shot and few-shot settings; (iv) systematic evaluation of seven open weight small and medium sized LLMs on formally verified proof generation tasks.

The remainder of this paper is organized as follows. Section 2 introduces the theoretical foundations, Section 3 details the experimental methodology adopted, and Section 4 presents the results. Finally, Section 5 concludes and outlines future work.

2. Theoretical Foundations

2.1. Propositional Logic

Propositional Logic is a formal language focused on representing and analyzing propositions and their logical relations. It is based on propositions or declarative sentences for which one can argue their truth or false [Huth and Ryan 2004]. Declarative sentences considered atomic, such as "*it rained*." are called propositional atoms. The set of these atoms, together with connectives (negation $\neg$, conjunction $\wedge$, disjunction $\vee$ and implication $\rightarrow$) and the symbols '(' and ')', constitute the alphabet of Propositional Logic. This language can be defined by a grammar in *Backus Naur Form* (BNF), where $\bot$ represents contradiction and P any propositional atom. Furthermore, each occurrence φ to the right of '::=' represents any formula already constructed, as shown below [Huth and Ryan 2004]:

$$\varphi ::= \bot \mid P \mid (\neg\varphi) \mid (\varphi \wedge \varphi) \mid (\varphi \vee \varphi) \mid (\varphi \rightarrow \varphi)$$

2.2. First Order Logic (FOL)

First-Order Logic (FOL) [Huth and Ryan 2004] extends Propositional Logic by allowing the representation of properties and relationships among objects in a more expressive way. FOL language introduces predicates, constants, variables, quantifiers and functional symbols. Predicates represent properties or relationships among elements of the universe of discourse and the quantifiers (universal $\forall$ and existential $\exists$) allow expressing generalizations and existential statements. FOL language is composed of terms and formulas. Terms denote objects in the domain and can be variables, constants or functions applied to other terms. Formulas express statements that are semantically evaluate as true or false. Formally, its grammar can be defined in BNF as (t are terms and φ are formulas):

$$t ::= x \mid c \mid f(t_1, \ldots, t_n)$$

$$\varphi ::= \bot \mid P(t_1, \ldots, t_n) \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \rightarrow \varphi \mid \forall x\varphi \mid \exists x\varphi$$

In formulas, variables can occur either free or bound, depending on whether they are within the scope of a quantifier. Furthermore, term substitution, denoted by $\varphi[x \leftarrow t]$, consists of replacing the free occurrences of a variable by a term [Huth and Ryan 2004].

2.3. Natural Deduction

Natural Deduction is a deductive system in which proofs are constructed through the structured application of introduction and elimination rules for logical connectives, aiming to establish that $\Gamma \vdash \varphi$, where $\Gamma = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ is a finite set of formulas representing the premises, and φ is the conclusion to be derived. The symbol $\vdash$ denotes syntactic entailment, indicating that φ can be formally derived from the set of premises Γ within the system. The work by [Pelletier 2000] identifies the Fitch format as the most widely adopted in textbooks, characterized by the linear representation of inferences and the explicit delimitation of subproofs through boxes.

In this system, each line of the proof contains a formula and the corresponding justification, which can be a premise or the application of a deductive rule. Subproofs are introduced to represent temporary hypotheses, whose scope is controlled by the opening and closing of boxes, allowing the local use of the assumption within the derivation. The hierarchy of boxes, potentially nested or interleaved, determines the access restrictions to formulas already introduced: a formula may only be referenced while it remains within the scope of the subproof in which it was derived.

Each logical connective is associated with introduction and elimination rules, such as: conjunction introduction and elimination ($\wedge i$, $\wedge e$), disjunction introduction and elimination ($\vee i$, $\vee e$), implication introduction and elimination ($\rightarrow i$, $\rightarrow e$), and negation introduction and elimination ($\neg i$, $\neg e$). Additional rules such as contradiction elimination ($\bot e$), *reductio ad absurdum* (raa) and copy are also part of the basic rule set of this deductive system. In FOL, these rules are extended with quantifier rules, namely universal introduction and elimination ($\forall i$, $\forall e$) and existential introduction and elimination ($\exists i$, $\exists e$).

2.4. Analytical Tableau

Analytical Tableau is a deductive system that operates through the systematic decomposition of formulas, constructing a tree structure that represents the possible interpretations

of a set of premises. This process allows us to check, in a general way, whether a conclusion follows from a set of premises or whether there is a situation in which the premises are true and the conclusion is false. The proof is constructed by refutation: in order to establish that $\Gamma \vdash \varphi$, we initially assume that all formulas in Γ are true and that φ is false. Starting from this assumption, we seek to find a contradiction, which would prove that our assumption about φ being false is incorrect, thereby confirming $\Gamma \vdash \varphi$. If no contradiction is found, it indicates that in at least one situation Γ is true and φ is false, producing a counterexample [Silva et al. 2017].

Each node in the tableau is labeled with T or F, indicating that the formula is assumed to be true or false under the interpretation being constructed.The decomposition rules follow the α/β classification: α-rules expand formulas within the same branch (e.g., negation, conjunction true, disjunction false, implication false), while β-rules introduce branching (e.g., conjunction false, disjunction true, implication true). The process is *refutation-complete*, meaning every compound formula is eventually decomposed.

In FOL, additional rules handle quantifiers, namely universal and existential rules in both true and false cases ($\forall T$, $\forall F$, $\exists T$, $\exists F$), subject to standard restrictions on term substitution and the introduction of fresh variables. A branch is saturated when no more rules are applicable to the formulas. Closure occurs when a branch contains a contradictory pair T φ and F φ. If all branches close, the tableau is closed and the derivation $\Gamma \vdash \varphi$ is established. Conversely, the existence of a saturated and open branch is an interpretation that satisfies Γ but falsifies φ, serving as a counterexample [Smullyan 1995].

2.4.1. Proof Assistants for Natural Deduction and Analytical Tableau

Both proof assistants were designed to support undergraduate students in verifying reasoning steps. They are inspired by the Fitch-style notation used in formal proofs as it's input language and cover both propositional and first-order logic, with only minor syntactic differences.

Figure 1 presents an example proof constructed in both systems of the same problem, illustrating how proofs are constructed in each system.

Natural Deduction Proof	**Analytical Tableau Proof**
1. $Ax(B(x) \mid B(x))$ pre	1. $T\, Ax(B(x) \mid B(x))$ pre
2. $\{\, a \,\}$	2. $F\, Ax\, B(x)$ conclusion
3. $B(a) \mid B(a)$ Ae 1	3. $F\, B(a)$ 2
4. $\{\, B(a) \,\}$ hip	4. $T\, (B(a) \mid B(a))$ 1
5. $B(a)$ copie 4	5. $\{\, T\, B(a) \,\}$ 4
6. $\{\, B(a) \,\}$ hip	6. $\oplus$ 3,5
7. $B(a)$ copie 6	7. $\{\, T\, B(a) \,\}$ 4
8. $B(a)$ $\mid e$ 3,4–5,6–7	8. $\oplus$ 3,7
9. $Ax\, B(x)$ Ai 2–8	

Figure 1. Proof of $(Ax(B(x) \mid B(x))) \vdash Ax\, B(x)$ in Natural Deduction and Analytical Tableau.

2.5. Large Language Models (LLMs)

Large Language Models (LLMs) constitute one of the main recent developments in Artificial Intelligence (AI), particularly in the field of Natural Language Processing (NLP). These models are based on deep neural networks trained on large volumes of textual data, enabling the contextualized and coherent understanding and generation of language [Brown et al. 2020]. Modern generative LLMs predominantly employ an autoregressive decoder-only transformer architecture, leveraging attention mechanisms [Vaswani et al. 2017] to evaluate complex linguistic dependencies without relying on a separate encoding module. The foundational capability of these models is established through an extensive pre-training phase focused entirely on next-token prediction, where the network optimizes its parameters by iteratively calculating the probability distribution of the subsequent discrete token given the preceding text sequence.

3. Methodology

This study investigates the performance of seven LLMs in constructing formal proofs in Propositional Logic and FOL, in the Natural Deduction and Analytical Tableau systems. Figure 2 provides an overview of the methodological steps adopted in this investigation. All problems, prompts, and evaluation scripts are publicly available in the GitHub repository associated with this submission. [1]

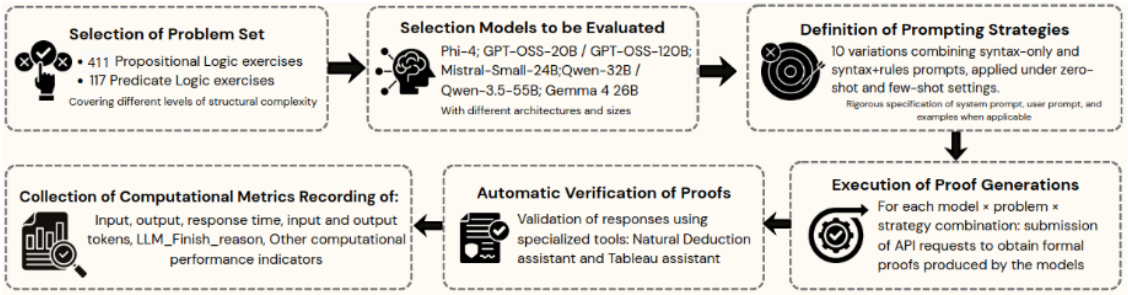


Figure 2. Methodological workflow adopted in this study.

3.1. Problem Set

A set of 411 Propositional Logic problems and 117 First-Order Logic (FOL) problems was used. Among the logical problems, 274 propositional instances were newly created for this study, while the remaining problems were obtained from previously developed educational materials used in undergraduate logic courses. The same set of problems was evaluated under both Natural Deduction and Analytic Tableaux.

Table 1 summarizes the distribution of logical operators across the dataset. The reported statistics show substantial variation in the occurrence of negations, conditionals, conjunctions, and disjunctions, with all operator types ranging from absent in some problems to multiple occurrences in others. Consequently, the dataset created for this study covers diverse reasoning patterns and proof structures, requiring models to handle different combinations of inference rules rather than repeatedly applying the same deductive strategy.

[1] anonymous.4open.science/r/FinalResults-5FF4

Table 1. Problems Distribution

	Propositional Problems				First-Order Problems			
	mean	median	min	max	mean	median	min	max
N° of negations in the problem	1	0	0	4	0.96	0	0	4
N° of conditionals in the problem	1.69	2	0	6	1.09	1	0	4
N° of disjunctions in the problem	0.74	0	0	6	0.29	0	0	2
N° of conjunctions in the problem	1.09	1	0	11	0.47	0	0	2

3.2. Evaluated Models

This study evaluates seven open weight LLMs spanning different architectures, parameter scales, and training paradigms in order to analyze their impact on formal reasoning performance. The evaluated models are:

- **Phi-4** (15B) [Abdin et al. 2024],
- **GPT-OSS-20B** and **GPT-OSS-120B** [Agarwal et al. 2025],
- **Mistral-Small-3.2-24B-Instruct-2506-GGUF** [Mistral AI 2025],
- **Qwen3-32B** [Yang et al. 2025] and **Qwen3.5-35B-A3B** [Qwen Team 2026],
- **Gemma 4 26B-A4B-IT** [Abd and et al. 2026].

These models represent recent state-of-the-art open-weight systems and provide a diverse set of reasoning capabilities, enabling the analysis of formal proof generation across different model families and parameter scales.

3.3. Prompting Strategies

All interactions with the models were performed through API calls using standardized prompts designed to elicit formally structured proofs. In order to ensure reproducibility, all prompts, examples, and evaluation scripts are publicly available in our GitHub repository.

Each prompt consisted of three components: (i) a *system prompt*, which specified the deductive system and output requirements; (ii) when applicable, a set of input–output examples in the few-shot setting; and (iii) a *user prompt*, containing the logical problem to be solved. We evaluated ten prompting configurations organized along two dimensions: prompt complexity and number of in-context examples:

- **Syntax-Only Zero-Shot (S0):** high-level description of the proof format, without examples;
- **Syntax-Plus-Rules Zero-Shot (C0):** complete specification of the deductive system, including inference rules and syntactic constraints, without examples;
- **Syntax-Only Few-Shot (S1, S2, S3, S5):** syntax description together with 1, 2, 3, or 5 solved proof examples;
- **Syntax-Plus-Rules Few-Shot (C1, C2, C3, C5):** complete deductive-system specification together with 1, 2, 3, or 5 solved proof examples.

4. Results

This section presents the experimental results in tabular form and discusses the performance of the evaluated models. Formal correctness was assessed exclusively in terms of accuracy through automatic verification using specialized proof assistants. It's also of note that the column "Best Prompt Diff." denotes the difference between the best performance obtained using complex prompts and the best performance obtained using simple prompts, expressed in percentage points (p.p.).

4.1. Natural Deduction Results

Table 2. Propositional Logic Natural Deduction Results

Model	Simple Prompts					Complex Prompts					Best Prompt Diff.
	S0	S1	S2	S3	S5	C0	C1	C2	C3	C5	
Phi-4	0.99%	8.37%	13.55%	**16.50%**	15.27%	1.48%	16.26%	**20.94%**	18.23%	17.24%	**+4.44 p.p.**
Gpt-20B	17.00%	23.40%	**29.80%**	29.31%	28.33%	24.63%	40.39%	45.32%	**45.57%**	40.39%	**+16.27 p.p.**
Mistral-24B	0.74%	15.02%	20.69%	**21.18%**	20.69%	3.20%	21.92%	22.17%	27.09%	**27.34%**	**+6.16 p.p.**
Qwen3-32B	46.31%	51.72%	**60.59%**	58.62%	54.19%	50.00%	61.33%	**67.24%**	65.27%	63.55%	**+6.65 p.p.**
GPT-120B	45.81%	53.69%	**62.81%**	57.88%	57.14%	62.32%	67.73%	**71.43%**	67.00%	65.52%	**+8.62 p.p.**
Qwen3.5-35B	4.68%	17.00%	**21.43%**	15.02%	20.44%	10.34%	13.05%	**19.46%**	15.27%	18.47%	**-1.97 p.p.**
Gemma-4-26B	18.97%	38.42%	**39.66%**	33.00%	35.47%	45.07%	**50.99%**	49.01%	47.04%	47.54%	**+11.33 p.p.**

Table 3. FOL Natural Deduction Results

Model	Simple Prompts					Complex Prompts					Best Prompt Diff.
	S0	S1	S2	S3	S5	C0	C1	C2	C3	C5	
Phi-4	0.00%	1.79%	**12.50%**	1.79%	4.46%	0.89%	8.04%	**18.75%**	2.68%	12.50%	**+6.25 p.p.**
Gpt-20B	1.79%	0.89%	**8.04%**	4.46%	4.46%	8.93%	1.79%	**9.82%**	8.93%	5.36%	**+1.78 p.p.**
Mistral-24B	0.00%	1.79%	12.50%	**17.86%**	13.39%	1.79%	9.82%	19.64%	17.86%	**22.32%**	**+4.46 p.p.**
Gemma-4-26B	12.50%	13.39%	17.86%	10.71%	**19.64%**	12.50%	16.07%	**16.96%**	9.82%	8.04%	**-2.68 p.p.**
Qwen3-32B	3.57%	12.50%	21.43%	23.21%	**32.14%**	10.71%	18.75%	**31.25%**	25.00%	25.89%	**-0.89 p.p.**
Qwen3.5-35B	0.00%	0.00%	16.96%	**19.64%**	8.04%	3.57%	1.79%	17.86%	**25.00%**	9.82%	**+5.36 p.p.**
Gpt-120B	2.68%	9.82%	8.93%	15.18%	**22.32%**	18.75%	25.00%	**33.04%**	32.14%	31.25%	**+10.72 p.p.**

Tables 2 and 3 report results for propositional logic and FOL Natural Deduction under different prompting strategies. The best propositional performance is achieved by GPT-120B **(71.43%)**, followed by Qwen3-32B at **34.98%**, representing a substantial performance gap (Table 2). In FOL (Table 3), GPT-120B again achieves the highest accuracy (33.04%). However, Qwen3-32B performs comparably, with a difference of only 0.9 percentage points. The substantial performance differences between models with markedly different parameter counts suggest that factors beyond model size, such as architecture and training, may play an important role in formal reasoning performance.

Despite these results, overall performance remains limited, indicating substantial room for improvement in formal proof generation. As expected, accuracy drops sharply from propositional to first-order logic across all models, reflecting the added complexity of quantifiers and variable binding.

Few-shot prompting consistently outperforms zero-shot across all settings, highlighting the importance of demonstration-based guidance. The few-shot configurations are designed to control rule coverage: S1 provides partial exposure, S2 ensures that each inference rule appears at least once, and higher-shot settings reinforce rules through repeated usage. For simple prompts, which is composed of format only instructions, the best propositional results are typically obtained at S2, indicating that minimal but complete rule coverage is sufficient. In contrast, first-order performance often peaks at higher-shot settings (e.g., S5), suggesting that repeated exposure is necessary for more complex reasoning.

For complex prompts, which consists of the simple prompt plus explanations and schemes of all rules in the deduction system, the optimal number of examples varies across models, indicating no universally optimal configuration. Since these prompts explicitly explain each rule and illustrate its application scheme, they provide richer guidance. This richer context likely explains why complex prompts consistently outperform simple ones, although the magnitude of improvement differs by model.

4.2. Tableau Results

Table 4. Propositional Logic Tableau Results

Model	Simple Prompts					Complex Prompts					Best Prompt Diff.
	S0	S1	S2	S3	S5	C0	C1	C2	C3	C5	
Phi-4	0.00%	0.99%	1.97%	0.74%	**4.43%**	0.74%	2.71%	4.43%	**6.90%**	6.65%	**+2.47 p.p.**
Gpt-20B	1.48%	8.87%	14.29%	14.53%	**15.27%**	17.49%	21.18%	23.15%	23.95%	**29.31%**	**+14.04 p.p.**
Mistral-24B	0.00%	0.74%	1.72%	4.19%	**5.42%**	0.99%	3.20%	7.39%	**9.85%**	9.61%	**+4.43 p.p.**
Gemma-4-26B	0.00%	5.17%	5.67%	9.61%	**10.10%**	1.72%	4.43%	5.42%	**9.36%**	8.87%	**-0.74 p.p.**
Qwen3-32B	4.42%	3.93%	5.69%	7.14%	**13.55%**	18.97%	12.04%	21.43%	20.94%	**34.98%**	**+21.43 p.p.**
Qwen3.5-35B	0.25%	0.99%	5.17%	5.91%	**9.85%**	3.94%	3.45%	6.90%	9.11%	**13.79%**	**+3.94 p.p.**
Gpt-120B	15.02%	22.66%	21.18%	22.41%	**55.17%**	47.29%	47.29%	41.38%	46.06%	**60.10%**	**+4.93 p.p.**

Table 5. FOL Tableau Results

Model	Simple Prompts					Complex Prompts					Best Prompt Diff.
	S0	S1	S2	S3	S5	C0	C1	C2	C3	C5	
Phi-4	0.00%	2.65%	1.77%	6.19%	**15.93%**	0.00%	1.77%	7.08%	9.73%	**16.81%**	**+0.88 p.p.**
Gpt-20B	0.00%	3.54%	7.08%	6.19%	**7.08%**	30.09%	33.63%	32.74%	35.40%	**35.40%**	**+28.32 p.p.**
Mistral-24B	0.00%	1.77%	1.77%	10.62%	**14.16%**	0.00%	10.62%	7.96%	**13.27%**	4.42%	**-0.89 p.p.**
Gemma-4-26B	0.88%	1.77%	6.19%	8.18%	**15.04%**	2.65%	15.04%	20.35%	**25.66%**	19.47%	**+10.62 p.p.**
Qwen3-32B	0.00%	**16.81%**	6.19%	8.85%	10.62%	**25.66%**	**25.66%**	21.24%	21.24%	22.12%	**+8.85 p.p.**
Qwen3.5-35B	0.88%	7.08%	11.50%	**19.47%**	15.04%	6.19%	8.85%	12.39%	15.04%	**22.12%**	**+2.65 p.p.**
Gpt-oss-120B	15.93%	31.86%	34.51%	**41.59%**	34.51%	80.53%	77.88%	76.11%	80.53%	**85.84%**	**+44.25 p.p.**

In the propositional setting, Gpt-oss-120b achieves the strongest performance (**60.10%**), substantially outperforming all other models. The second-best model, Qwen3-32b (34.98%), followed by Gpt-oss-20b (29.31%), while the remaining models perform considerably worse.

Notably, this ranking does not align strictly with model size. For example, Qwen3.5-35b substantially underperforms Qwen3-32b despite having a larger parameter count. This suggests that performance cannot be explained by scale alone and indicates that factors such as model architecture and training may also play an important role in Tableau-style reasoning. Gpt-oss-120b, in particular, demonstrates a notable strength on this task. The same pattern is observed in the FOL setting (Table 5), where Gpt-oss-120b achieves **85.84%**, the highest score across all experiments.

As observed for Natural Deduction, both few-shot prompting and complex prompts consistently improve performance, further highlighting the importance of explicit guidance and structured demonstrations for formal reasoning tasks.

Interestingly, unlike Natural Deduction, FOL Tableau performance does not degrade relative to the propositional setting, instead, it improves across models, with the notable exception of Qwen3-32B. This result is unexpected given the additional complexity of quantifiers and variable instantiation.

One possible explanation is that the relative scarcity of such structured reasoning patterns in pretraining data may make models more responsive to explicit instructions and demonstrations. This could help explain the substantial gains observed with complex prompting, including an improvement of 25.74 percentage points for Gpt-oss-120b. However, this interpretation remains speculative and would require further investigation to determine the underlying cause of the observed performance gains.

5. Conclusion

This work investigated the capability of large language models to perform formal logical reasoning under strict deductive constraints, focusing on proof construction in Propositional and First-Order Logic within Natural Deduction and Analytical Tableaux systems. By relying on symbolic datasets with human-constructed and formally verified proofs, we eliminated the ambiguities inherent in natural language benchmarks and enabled a rigorous evaluation of logical correctness through automatic proof checking.

Our experiments covered a diverse set of open-weight small and medium-sized models, evaluated under zero-shot and few-shot prompting settings. The results consistently show that prompting strategies play a significant role in performance, with few-shot configurations generally improving proof quality, particularly when combined with more structured prompt formulations. However, these improvements are not sufficient to bridge the gap toward fully reliable formal reasoning, as even larger models continue to exhibit failures in maintaining global structural consistency and strict adherence to inference rules.

The findings also highlight that model scale alone does not guarantee robust logical reasoning performance, and that smaller models remain highly sensitive to prompt design and example selection. While open-weight models provide an important foundation for reproducible research and controlled experimentation, their current capabilities in formal proof construction remain limited.

Finally, this study reinforces the importance of evaluating reasoning systems through formal, verifiable frameworks rather than natural language proxies. By introducing symbolic datasets spanning both constructive (Natural Deduction) and decision-based (Tableau) paradigms, along with propositional and first-order logic, this work contributes a more comprehensive benchmark for assessing logical reasoning in LLMs.

Future work includes investigating advanced prompt optimization strategies such as GEPA, exploring fine-tuning approaches for open-weight LLMs, and constraining LLM outputs using formal grammars. After all, the individual and combined assessment of these approaches will enable a more detailed understanding of the different factors affecting proof generation performance.

References

Abd, S. E. and et al. (2026). Gemma 4 technical report. *CoRR*, abs/2607.02770.

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. (2025). gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Arkoudas, K. and Batzoglou, S. (2025). Stress-testing the reasoning competence of language models with formal proofs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12376–12394.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. volume abs/2110.14168.

Creswell, A., Shanahan, M., and Higgins, I. (2023). Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

et al., S. H. (2024a). FOLIO: natural language reasoning with first-order logic. In Al-Onaizan, Y., Bansal, M., and Chen, Y., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, Florida, USA, November 12–16, 2024*, pages 22017–22031. Association for Computational Linguistics.

et al., S. H. (2024b). P-FOLIO: evaluating and improving logical reasoning with abundant human-written reasoning chains. In Al-Onaizan, Y., Bansal, M., and Chen, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12–16, 2024*, volume 2024 of *Findings of the Association for Computational Linguistics*, pages 16553–16565. Association for Computational Linguistics.

Huth, M. and Ryan, M. D. (2004). *Logic in computer science - modelling and reasoning about systems (2. ed.)*. Cambridge University Press.

Li, X., He, S., Lei, F., JunYang, J., Su, T., Liu, K., and Zhao, J. (2024a). Teaching small language models to reason for knowledge-intensive multi-hop question answering. ACL 2024:7804–7816.

Li, X., He, S., Wu, J., Yang, Z., Xu, Y., jun Jun, Y., Liu, H., Liu, K., and Zhao, J. (2024b). Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts. pages 11475–11485.

Lin, B. Y., Le Bras, R., Richardson, K., Sabharwal, A., Poovendran, R., Clark, P., and Choi, Y. (2025). Zebralogic: On the scaling limits of llms for logical reasoning. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37889–37905. PMLR.

Lu, Z., Zhou, A., Ren, H., Wang, K., Shi, W., Pan, J., Zhan, M., and Li, H. (2024). Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

McGinness, L. and Baumgartner, P. (2024). Automated theorem provers help improve large language model reasoning.

Mistral AI (2025). Model Card: Mistral Small 3.2 24B Instruct 2506. GGUF model repository by LM Studio Community. Accessed: 2026-05-01.

Mitra, A., Corro, L. D., Mahajan, S., Codas, A., Simões, C., Agrawal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., Palangi, H., Zheng, G., Rosset, C., Khanpour, H., and Awadallah, A. (2023). Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045.

Pelletier, F. J. (2000). A history of natural deduction and elementary logic textbooks. *Logical consequence: Rival approaches*, 1:105–138.

Qwen Team (2026). Qwen3.5-35B-A3B. GGUF model repository by LM Studio Community.

Saparov, A. and He, H. (2023). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shridhar, K., Stolfo, A., and Sachan, M. (2023). Distilling reasoning capabilities into smaller language models. ACL 2023:7059–7073.

Silva, F. S. C. d., Finger, M., and Melo, A. C. V. d. (2017). *Lógica para Computação*. Cengage Learning, São Paulo, 1 edition.

Smullyan, R. M. (1995). *First-Order Logic*. Dover Publications.

Sprague, Z. R., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. (2024). Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Tafjord, O. et al. (2021). Proofwriter: Generating implications, proofs, and abductive reasoning. In *ACL*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, P., Wang, Z., Li, Z., Gao, Y., Yin, B., and Ren, X. (2023). Scott: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Wang, S., Wei, Z., Choi, Y., and Ren, X. (2024). Can llms reason with rules? logic scaffolding for stress-testing and improving llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7523–7543, Bangkok, Thailand. Association for Computational Linguistics.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.