Immediate Inference: The Missing Foundation in Large Language Model Logical Reasoning

Sihang Jiang[1], **Zhiyu Lu**[1], **Keyi Wang**[1], **Jiaqing Liang**[2,*], **Yanghua Xiao**[1,*], **Xiaojun Meng**[3], **Jiansheng Wei**[3]
[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]School of Data Science, Fudan University
[3]Huawei Large Model Data Technology Lab
{jiangsihang, liangjiaqing, shawyh}@fudan.edu.cn
{luzy25, wangkeyi25}@m.fudan.edu.cn
{xiaojun.meng, weijiansheng}@huawei.com

Abstract

While extensive research has evaluated LLMs on complex reasoning tasks, the foundational building blocks of logical reasoning remain underexplored. We introduce IIBench, a benchmark evaluating immediate inference (elementary operations over categorical propositions). Our evaluation reveals that even SoTA models exhibit systematic deficiencies in immediate inference, and establishes immediate inference as foundational: it mediates approximately 40% of the effect on syllogistic reasoning, with near-perfect correlation ($\rho = 0.98$) across reasoning benchmarks. Our analysis reveals that models lack robust operator grounding, oscillating between structural reasoning and surface pattern matching with inconsistent handling of quantifiers and negation.[1]

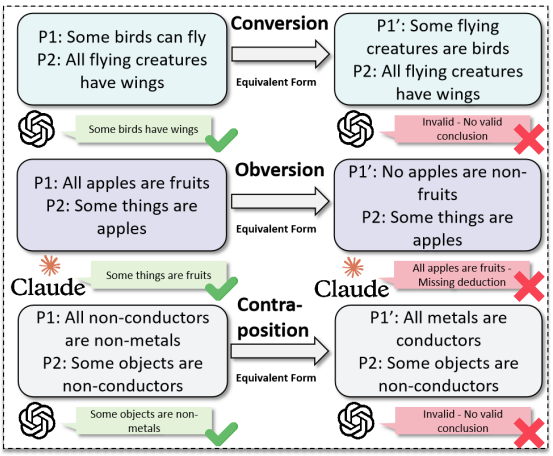


Figure 1: Immediate inference failures cascade into syllogistic reasoning errors.

1 Introduction

Despite remarkable performance on various reasoning benchmarks (Wei et al., 2022; Cobbe et al., 2021; Lewkowycz et al., 2022), large language models (LLMs) exhibit fragile and inconsistent logical reasoning abilities (Sap et al., 2022; Srivastava et al., 2022). Models that correctly solve a reasoning problem often fail on logically equivalent reformulations, misinterpret quantifiers, or display high sensitivity to superficial variations in problem presentation (Kassner and Schütze, 2020; Elazar et al., 2021; Ozeki et al., 2024; Zong and Lin, 2024). As illustrated in Figure 1, model performance drops sharply even though the underlying logical structure remains invariant. These failures highlight the urgent need to identify the root causes of reasoning fragility and develop principled approaches to enhance reasoning reliability.

Recent work on LLM logical reasoning has progressed from task-level benchmarks (Ontañón et al., 2022; Han et al., 2024) to fine-grained skill decomposition (Wan et al., 2024; Pan et al., 2023) and consistency probing at the propositional level (Ghosh et al., 2024). These efforts converge on persistent fragility: models show limited transfer, surface-form sensitivity, and inconsistent behavior on equivalent formulations despite extensive training (Berglund et al., 2023; McCoy et al., 2023). Existing approaches predominantly adopt a task-centric, top-down perspective, which evaluate performance on target reasoning tasks and decomposing failures into atomic skills to be learned through examples (Wan et al., 2024; Olausson et al., 2023).

Motivated by prerequisite learning theory in educational psychology, we ask **whether logical reasoning possesses core foundational capabilities?** Studies in mathematics education and cognitive science have demonstrated that complex cognitive abilities depend on mastering prerequisite skills (Fodor and Pylyshyn, 1988; Ausubel et al., 1978). For instance, algebraic reasoning requires prior fluency in arithmetic operations (Siegler, 1998). If logical reasoning follows this pattern, identifying and evaluating its foundational building blocks becomes essential for understanding and

[1]Resource of this paper can be found at https://github.com/michaellu5475/IIBench.git.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17766–17799

ultimately improving LLM reasoning capabilities.

We argue that **immediate inference (II) represents a systematically neglected foundation capability**. II operations (e.g., conversion, obversion and contraposition, as shown in Figure 1, reformulate categorical propositions without altering their logical content. From a formal logic perspective, these operations serve as compositional building blocks for syllogistic reasoning, analogical reasoning, and other inference types that require statement reformulation (Copi et al., 2016; DeLancey, 2017). Furthermore, II establishes understanding of quantifiers, which serves not only as the foundation for formal reasoning but also as a bridge connecting formal and natural language reasoning (Richardson et al., 2020; Geiger et al., 2020). Understanding these elementary transformations is essential for reliable complex reasoning, yet their mastery in current LLMs remains largely unexamined.

In this paper, we construct the first systematic **Bench**mark for evaluating **I**mmediate **I**nference in LLMs (i.e., `IIBench`), encompassing both transformation operations and opposition relations to assess models' capabilities in executing and understanding II[2], which contains 5,284 instances. Our empirical study finds that **even SoTA models exhibit systematic deficiencies in II**. Causal mediation analysis establishes II as foundational to downstream reasoning, with natural indirect effects approximately 40% for syllogistic reasoning, and strong correlations ($\rho = 0.98$) with formal reasoning benchmarks including multi-step deduction, first-order logic, and natural language inference. Moreover, our analysis reveals systematic instability: models apply inconsistent reasoning standards, oscillating between structural reasoning and surface pattern matching, fundamentally limiting reasoning reliability. Further experiments validate the potential for improving downstream reasoning by enhancing II capabilities.

Our contributions are summarized as follows. (1) To the best of our knowledge, we are the first to identify II capability as a foundational prerequisite for logical reasoning. (2) We construct the first systematic benchmark (`IIBench`) for evaluating II, revealing systematic deficiencies in SoTA LLMs. (3) We propose a causal analysis framework establishing both the causal and generalization properties of II as a foundational prerequisite, while revealing that models lack robust operator grounding and oscillate between structural reasoning and surface pattern matching. (4) We validate that training can improve downstream reasoning performance by enhancing II, while exposing the limitations of current paradigms in resolving the trade-off between structural and context-sensitive reasoning.

[2]II consists of two core components: transformation operations (conversion, obversion, contraposition) and opposition relations (contradictory, contrary, subcontrary, subaltern). Detailed definitions are provided in Section 3 and Appendix C.

2 Related Work

Benchmarking Logical Reasoning. Extensive efforts have been devoted to evaluating reasoning capabilities in language models, spanning multi-step syllogistic reasoning (Liu et al., 2021b; Yu et al., 2020; Zhong et al., 2021; Ozeki et al., 2024), deductive inference (Tafjord et al., 2021; Han et al., 2024), natural language inference (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020), compositional generalization (Lake and Baroni, 2018; Kim and Linzen, 2020; Keysers et al., 2019), and unified logical reasoning across propositional, first-order, and non-monotonic logic (Parmar et al., 2024). Despite this breadth, existing benchmarks overlook the foundational role of II itself.

Analyzing Reasoning Failures. Prior analyses have examined reasoning failures through diverse lenses: shortcut learning and spurious correlations expose reliance on surface patterns (McCoy et al., 2019; Gururangan et al., 2018); semantic paraphrase studies reveal inconsistent behavior across meaning-preserving reformulations (Ribeiro et al., 2019); adversarial perturbations demonstrate brittleness to superficial input variations (Jia and Liang, 2017; Ribeiro et al., 2018); investigations into quantifier understanding and monotonicity reasoning uncover systematic deficiencies in handling logical operators (Yanaka et al., 2019; Richardson et al., 2020). However, these efforts neglect causal links between logical structural transformations and downstream failures.

Improving Reasoning Capabilities. A variety of methods have been proposed to enhance reasoning, including chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022), reasoning-focused fine-tuning (Zelikman et al., 2022; Ho et al., 2023), neuro-symbolic integration (Olausson et al., 2023), self-consistency decoding (Wang et al., 2022), process reward models for step-level supervision (Lightman et al., 2023; Wang et al., 2024), test-

time compute scaling (Snell et al., 2024), and reasoning-specialized architectures (OpenAI, 2024; Guo et al., 2025). These approaches presuppose robust foundations, yet improvements may be constrained when foundational capabilities are brittle.

Detailed analysis and extended discussion of related work are provided in Appendix A.

3 Formal Definition

3.1 Definition and Foundational Role

Immediate inference refers to logical operations that derive a new categorical proposition directly from a single premise without requiring additional premises (Smith, 1989; Copi et al., 2016). A categorical proposition expresses a relationship between two classes using quantifiers, taking one of four standard forms: **A-type** (universal affirmative), **E-type** (universal negative), **I-type** (particular affirmative), and **O-type** (particular negative). The specific structures and examples of these forms are shown in Table 1.

II occupies a foundational position in evaluating and enhancing language model reasoning capabilities. As the simplest type of reasoning in formal logic (Kneale and Kneale, 1984), II is a necessary condition for more complex reasoning. If a model performs poorly on II, it is unlikely to succeed genuinely in multi-premise reasoning tasks such as syllogisms (Copi et al., 2016). The core of II is the understanding and manipulation of quantifiers (universal vs. particular, affirmative vs. negative), which is not only fundamental to formal reasoning but also closely related to natural language inference (Barwise and Cooper, 1981; Montague, 1973). As the intersection of formal reasoning and natural language inference, II serves as a critical probe for diagnosing model cognitive architectures. Furthermore, since transformation rules are determined entirely by logical form and are independent of specific content (Copi et al., 2016; Tarski, 1956), models that master these rules naturally acquire cross-content systematic generalization capabilities (Fodor and Pylyshyn, 1988; Chomsky, 2014).

3.2 Two Core Components of II

In this work, we adopt a broad conception of II that encompasses two fundamental components: *transformation operations* and *opposition relations*, which respectively embody the capabilities of propositional transformation and propositional understanding. Detailed discussions and examples are provided in Appendices B and C.

Type	Logical Form	Example
A (Universal Affirmative)	All S are P	All humans are mortal
E (Universal Negative)	No S are P	No birds are mammals
I (Particular Affirmative)	Some S are P	Some dogs are brown
O (Particular Negative)	Some S are not P	Some cats are not black

Table 1: 4 standard forms of categorical propositions.

Transformation Operations Transformation operations systematically manipulate the structure of categorical propositions while preserving or establishing logical relationships. Three primary transformation operations form the foundation of II: **conversion**, **obversion**, and **contraposition**. Conversion exchanges the subject and predicate terms of a proposition. Obversion changes the quality of a proposition (from affirmative to negative or vice versa) while replacing the predicate with its complement. Contraposition exchanges the subject with the complement of the predicate, and the predicate with the complement of the subject. Transformation operations do not preserve logical equivalence for all proposition types. Validity patterns are summarized in Table 2, with detailed justifications provided in Appendix C.

Opposition Relations Opposition relations characterize the logical dependencies between propositions that share identical subject and predicate terms but differ in quantity, quality, or both. The traditional square of opposition establishes four fundamental logical relations: **contradictory**, **contrary**, **subcontrary**, and **subaltern**, which define patterns of logical compatibility and entailment. Contradictory opposition relates propositions that differ in both quantity and quality (A-O and E-I pairs). Contradictory propositions cannot both be true and cannot both be false; the truth of one necessarily implies the falsity of the other. Contrary opposition relates universal propositions of opposite quality (A-E pairs). Contrary propositions cannot both be true, though both may be false. Subcontrary opposition relates particular propositions of opposite quality (I-O pairs). Subcontrary propositions cannot both be false, though both may be true. Subaltern opposition relates propositions of the same quality but different quantity (A-I and E-O pairs). The truth of the universal proposition (subaltern) entails the truth of the particular proposition (subalternate), but not vice versa.

4 Benchmark

4.1 Design Principles

The construction of IIBench follows three core design principles to ensure rigorous and comprehensive evaluation of II capabilities in LLMs.

Principle 1: Balancing Understanding and Execution. II requires two distinct but complementary capabilities: the ability to correctly execute transformation operations (execution) and the ability to understand opposition relations (understanding). Therefore, we design evaluation data based on transformation operations and opposition relations to ensure comprehensive coverage of both capabilities.

Principle 2: Disentangling Semantic Influence. We aim to evaluate whether models can distinguish genuine logical understanding from semantic associations. To this end, we design three types of propositions: factual, counterfactual, and anonymous, which respectively contain content consistent with real-world knowledge, conflicting with real-world knowledge, and devoid of semantic information.

Principle 3: Prompt Robustness. To mitigate the impact of task comprehension and model stability on evaluation results, we design multiple prompting strategies ranging from zero-shot prompts to prompts with explicit definitions and demonstrations, systematically testing model performance under different prompting conditions.

4.2 Data Construction

4.2.1 Seed Acquisition

The core challenge in II lies in determining the subject term S and predicate term P. We extract `instanceof` and `subclassof` relations from existing knowledge graphs (e.g., Wikidata (Vrandecic and Krötzsch, 2014)) as the data source for S and P. For example, through the `subclassof` relation, we obtain that "mammal" is a subclass of "animal", enabling us to construct propositions such as "All mammals are animals". This approach allows us to construct factual data. To satisfy the design principle of disentangling semantic influence, we randomly sample from classes and entities to generate propositions that conflict with world knowledge but remain semantically coherent, constructing the counterfactual data. Meanwhile, we anonymize S and P by replacing them with symbols (e.g., X, Y, Z) or meaningless strings (e.g., xadf, bklm) to completely eliminate semantic information, thereby constructing the anonymous data.

4.2.2 Evaluation Tasks

Based on the acquired subject-predicate pairs (S, P), we design three evaluation tasks to systematically assess models' II capabilities across both execution and understanding dimensions.

Transformation Operation Task. Given a proposition $\varphi \in \{A, E, I, O\}$ and a transformation operation τ (i.e., conversion, obversion, contraposition), the model must judge whether the transformed proposition $\varphi' = \tau(\varphi)$ is correct.

Opposition Relations Task. Given a true proposition $\varphi \in \{A, E, I, O\}$ and another proposition $\psi \in \{A, E, I, O\}$ from the square of opposition, the model must judge the truth value of ψ based on the opposition relation.

Syllogistic Transformation Task. Given a valid syllogism $(\pi_1, \pi_2) \vdash \gamma$, where π_1, π_2 are premises and γ is the conclusion, we randomly select a premise $\pi_i \in \{\pi_1, \pi_2\}$ and apply a transformation operation τ to obtain $\pi_i' = \tau(\pi_i)$. The model must judge both the validity of the transformed syllogism $(\pi_1', \pi_2) \vdash \gamma$ or $(\pi_1, \pi_2') \vdash \gamma$ and the truth value of the conclusion γ.

4.2.3 Data Statistics and Quality

Detailed construction specifics are provided in Appendix D.

5 Empirical Study

5.1 Overall Performance

Experimental Setup. We evaluate 8 representative LLMs, including 4 closed-source models and 4 open-source models: Gemini-2.5-Pro, GPT-o3, GPT-4.1, Claude-3.5-Sonnet, Qwen-2.5-72B, Llama-3.1-70B, Qwen-2.5-7B, Llama-3.1-8B. We use accuracy as the evaluation metric across two dimensions. All experiments are conducted under zero-shot, factual data conditions. Results are shown in Tables 2 and 3, detailed examples and more analysis are in Appendix E.1.

No model achieves reliability standards. Logical reasoning demands 100% certainty, yet even the strongest models exhibit unpredictable failures. Gemini-2.5-Pro achieves near-perfect performance (99-100%) on most operations but catastrophically collapses to 10.0% on contraposition E→Inv. GPT-o3 achieves 97.9% (transformation operations) and 77.1% (opposition relations), similarly facing reliability deficiencies. Open-source models per-

Model	Conversion				Obversion				Contraposition				By Premise				Overall
	A→I	E→E	I→I	O→Inv	A→E	E→A	I→O	O→I	A→A	E→Inv	I→Inv	O→O	A	E	I	O	Avg
Closed-Source Models																	
Gemini-2.5-Pro	**99.5**	**100.0**	**100.0**	**100.0**	**99.3**	**100.0**	**100.0**	**100.0**	**99.3**	10.0	96.0	**99.3**	**99.4**	94.0	**99.7**	**99.6**	**98.4**
GPT-o3	96.8	99.5	**100.0**	**100.0**	99.0	99.0	99.3	**100.0**	96.8	**92.0**	**100.0**	94.5	97.3	**98.8**	**99.7**	96.7	97.9
GPT-4.1	2.0	99.0	99.7	99.0	88.7	81.7	86.3	27.3	71.0	6.0	**100.0**	12.7	53.5	85.8	94.4	25.7	60.6
Claude-3.5-Sonnet	26.5	85.6	97.7	53.0	0.0	51.0	68.3	82.0	77.2	0.0	2.0	38.0	43.7	66.0	79.5	52.7	57.4
Open-Source Models																	
Qwen-2.5-72B	0.0	75.0	94.7	0.0	65.7	79.0	62.0	41.3	30.3	2.0	0.0	10.8	29.0	71.7	75.2	18.9	43.8
Llama-3.1-70B	0.0	70.2	94.2	0.0	53.0	57.3	45.0	40.0	22.2	0.0	58.0	12.0	22.3	60.3	72.0	19.2	38.7
Qwen-2.5-7B	0.0	36.4	30.6	36.0	0.3	5.3	37.3	64.3	3.5	0.0	0.0	12.0	1.7	21.4	31.2	30.1	18.8
Llama-3.1-8B	0.0	3.3	48.5	0.0	0.0	56.3	12.0	6.0	21.7	0.0	0.0	11.0	9.9	24.4	30.6	8.4	16.4

Table 2: Transformation operation performance across proposition types. "Inv" denotes invalid transformation. Best results are in **bold**, second-best are underlined. **Even SoTA models exhibit systematic deficiencies in II.**

Model	Premise A			Premise E			Premise I			Premise O			Overall
	A→E	A→I	A→O	E→A	E→I	E→O	I→A	I→E	I→O	O→A	O→E	O→I	Avg
	Det.	*Det.*	*Det.*	*Det.*	*Det.*	*Det.*	*Und.*	*Det.*	*Und.*	*Det.*	*Und.*	*Und.*	
Closed-Source Models													
Gemini-2.5-Pro	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	99.0	**100.0**	**98.8**	**100.0**	**100.0**	**100.0**	**99.8**
GPT-o3	22.8	38.0	**100.0**	35.6	**100.0**	28.6	**100.0**	**100.0**	**98.8**	**100.0**	**100.0**	**100.0**	77.1
GPT-4.1	73.3	16.0	97.9	87.4	98.9	41.6	**100.0**	93.9	95.2	97.8	14.1	94.7	76.4
Claude-3.5-Sonnet	**100.0**	99.0	92.6	82.8	76.4	18.2	99.0	94.9	91.7	79.8	96.5	88.4	86.2
Open-Source Models													
Llama-3.1-70B	**100.0**	**100.0**	**100.0**	72.4	82.0	83.1	67.3	**100.0**	73.8	94.4	48.2	47.4	81.2
Llama-3.1-8B	96.0	97.0	94.7	40.2	88.8	37.7	0.0	**100.0**	0.0	98.9	0.0	0.0	55.6
Qwen-2.5-72B	99.0	99.0	**100.0**	83.9	89.9	41.6	99.0	**100.0**	89.3	88.8	20.0	83.2	84.2
Qwen-2.5-7B	96.0	84.0	56.4	73.6	60.7	50.6	76.2	99.0	57.1	89.9	8.2	86.3	71.1

Table 3: Truth value determination accuracy on the square of opposition. The logic type row indicates whether the relationship yields determinate truth/falsity (*Det.*) or is logically undetermined (*Und.*). Best results are in **bold**, second-best is underlined. **Except for Gemini-2.5-Pro, all models exhibit critical deficiencies in opposition relation reasoning.**

form significantly lower (transformation operations 16.4%-43.8%).

Applying unstable logical standards. Models exhibit task difficulty inversion. GPT-4.1 achieves only 16.0% on the basic A→I inference but reaches 97.9% on the more complex A→O relation, a gap of 81.9%. GPT-o3 achieves near 100% on undetermined types yet only 22.8% on determinate A→E relation. This inconsistency demonstrates that models oscillate between structural reasoning and surface pattern matching, applying inconsistent reasoning modes.

Lack of robust operator grounding for quantifiers. Models exhibit systematic difficulty with particular propositions. In opposition relations, inference from universal premises (A/E) significantly outperforms particular premises (I/O). Qwen-2.5-72B achieves 99.3% on premise A but only 64.0% on premise O. Models also display bidirectional judgment failures: committing to definite values in undetermined cases while abstaining from valid inferences in determinate cases (e.g., GPT-4.1 achieves only 16.0% on A→I).

5.2 Robustness Analysis

Experimental Setup. We test the effects of semantics and prompts on II across models[3]. The evaluation metric is accuracy gain relative to zero-shot baseline. Table 4 presents results aggregated across semantic conditions, Figure 2 shows gains under the anonymous condition; examples, prompts, and full stratification are in Appendix E.2.

Model scale modulates prompt effectiveness. Small models (<10B) benefit from prompting interventions but exhibit instability: contraposition improves by +31.2% with P3, while conversion degrades from +27.9% (P2) to +15.9% (P3). Medium models (∼70B) show the most substantial gains, achieving +73.6% total improvement on contrapo-

[3]We design a three-level prompt intervention strategy: P1 (zero-shot), P2 (w/ definition), and P3 (w/ definition & examples).

Task	Δ_{Def} (P2-P1)	Δ_{Example} (P3-P2)	Δ_{All} (P3-P1)
Conversion			
Small	0.279	-0.120	0.159
Medium	0.312	-0.014	0.298
Large	0.127	-0.005	0.123
Obversion			
Small	0.093	0.208	0.301
Medium	0.142	-0.022	0.120
Large	0.060	0.033	0.093
Contraposition			
Small	-0.005	0.317	0.312
Medium	0.481	0.255	0.736
Large	-0.048	-0.002	-0.050

Table 4: Prompt effectiveness across tasks and model scales, aggregated over all semantic conditions. **Model scale modulates prompt effectiveness.**

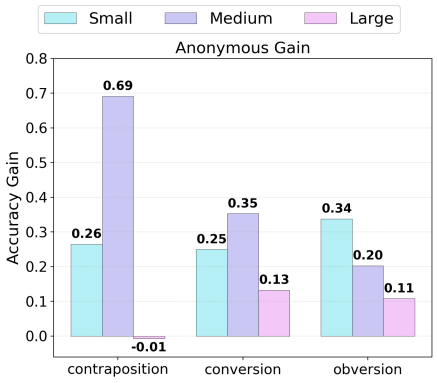


Figure 2: Prompt-induced accuracy gains stratified by anonymous condition. **Anonymization reveals models' reliance on semantic shortcuts.**

sition, with definitions contributing +48.1% and examples adding +25.5%. Large models (closed-source frontier models) maintain stable performance but show limited improvement, and notably exhibit negative interference on contraposition (total gain -5.0%), suggesting that additional definitions and examples may disrupt their established internal reasoning patterns.

Anonymization reveals reliance on semantic shortcuts. Removing semantic content amplifies prompt effects and exposes differential reliance on surface-level cues. Under counterfactual conditions, medium-scale models achieve +87.14% gain on contraposition, substantially higher than under factual (+64.29%) or anonymous (+69.12%) conditions. However, anonymization also exposes fragility in small models: they exhibit lower gains on conversion under anonymous conditions (+25.00%) compared to factual conditions (+32.86%), indicating that without semantic scaffolding, their insufficient mastery of structural reasoning leads to reduced benefits.

5.3 Failure Mechanism Analysis

Experimental Setup. We use a rule-based semantic parsing script to annotate error types. For understanding tasks, these include Hallucination (U→T, misjudging undetermined as True), Forced False (U→F, forcing judgment as False), Abstention (T/F→U, refusing to judge in determinate cases), and Polarity Flip (T↔F, reversing truth values). For execution tasks, these include Negation Error (missing or incorrect negation), Refusal/Format (refusing to execute or format errors), Over-Generation (forcing output for invalid transformations), and Quantifier Error (incorrect quantifiers). Figure 3 presents error distributions across models, detailed examples and prompts are in Appendix E.3.

Polarized risk preferences. Open-source models exhibit "action bias": forced binarization in understanding tasks (Forced False comprising 33.3%-63.3%). Closed-source models (especially GPT-4.1) exhibit "inaction bias": excessive abstention in understanding tasks (Abstention at 53.8%) and refusal to execute in execution tasks (Conversion Refusal reaching 94.8%). Both extreme biases lead to systematic failures across tasks.

Lack of operator grounding. LLMs fail to ground logical operators in their formal definitions. For quantifiers, models refuse to downgrade "All" to "Some", causing A→I determination failures and Quantifier Errors. In conversion tasks, open-source models exhibit Quantifier Error rates of 50.1%-72.8%. For negation, Negation Error dominates in obversion (72.5%-87.1%) and in contraposition (66.4%-84.6%), suggesting that negation is neither reliably preserved during generation nor accurately detected during evaluation, which indicates a dissociation between surface form and operator semantics.

Surface heuristics overriding structural rules. LLMs rely on lexical overlap heuristics that prioritize surface similarity over logical validity. In understanding tasks, models assume that propositions sharing identical subject-predicate pairs must share truth values, resulting in spurious equivalence judgments (Hallucination rates up to 38.6%, Polarity Flip up to 27.6%). In execution tasks, models generate outputs that maximize lexical overlap with the input, leading to improperly preserved quantifiers and omitted negations.

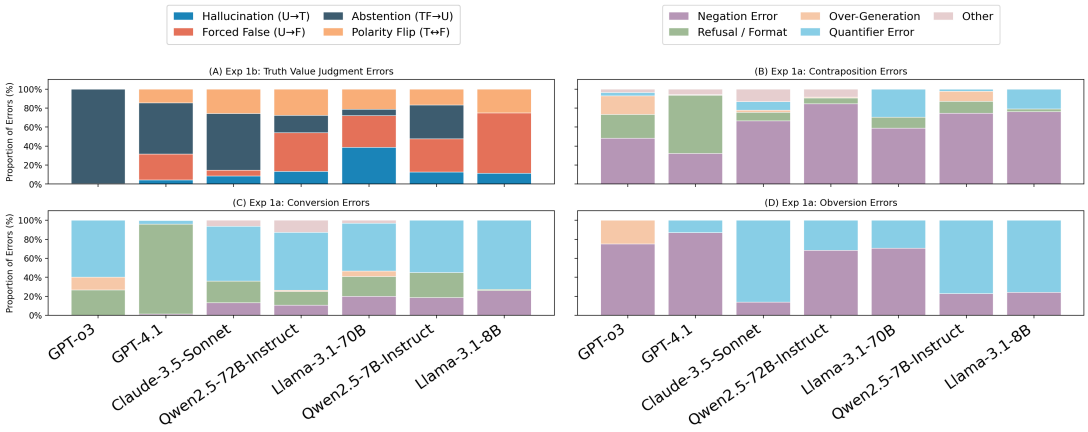


Figure 3: Error type distribution across understanding and execution dimensions. When multiple error types co-occur, we assign each sample to the highest-priority category following the order: Refusal/Format > Quantifier > Negation > Over-generation/Term mismatch. **Open-source models exhibit "action bias" while closed-source models show "inaction bias". Negation errors dominate across models.**

Path	Factual	Anonymous	Counterfactual
a_1: Scale → Execution	0.204	0.226	0.196
a_2: Scale → Understanding	0.049	0.050	0.043
b_1: Execution → Y	0.324	-0.058	0.054
b_2: Understanding → Y	0.145	**1.306**	**0.949**
c': Scale → Y	0.042	0.097	0.074

Table 5: Path analysis results across semantic conditions. Coefficients represent standardized path weights. **Models tend to bypass structural derivation when semantic cues are available.**

5.4 Causal Mechanism Analysis

Experimental Setup. We design two complementary experiments to analyze the causal influence of II on syllogistic reasoning. First, we examine the causal pathways from model scale to syllogistic reasoning performance (Y), mediated through two dimensions of II capability: execution and understanding. Analyses are conducted separately under factual, anonymous, and counterfactual conditions. Second, we apply logically equivalent transformations to syllogism premises and evaluate whether models can still derive valid conclusions. We then compute conditional probabilities to quantify the dependency between II (M) and syllogistic reasoning (Y): **TRC** ($P(Y = 1 \mid M = 1)$), **FI** ($P(Y = 0 \mid M = 0)$) and **SCR** ($P(M = 0 \mid Y = 1)$). Results are shown in Figure 4. Detailed settings and results are shown in Appendix E.4.

II as Foundation. The TRC heatmap shows that large models achieve higher causal sufficiency (TRC = 0.74–0.98 across operations), indicating that correct foundational logic leads to high success rates in syllogistic reasoning. Small models achieve TRC = 0.29–0.72, indicating lower efficiency in logical chain transmission. II capability also correlates strongly with downstream reasoning tasks (Spearman $\rho = 0.98$, $p < 0.0001$, details in Appendix F.1).

The presence of semantic context can obscure models' genuine logical reasoning capabilities. Under factual conditions, the understanding pathway shows no significant contribution to reasoning ($b_2 = 0.15$), suggesting that models may obtain answers through semantic shortcuts rather than structural derivation. Under anonymous/counterfactual conditions, this pathway becomes significant ($b_2 = 0.95$-1.31, $p < 0.05$). For conversion, small models exhibit higher SCR (0.37) than large models (0.20), indicating greater reliance on surface-level pathways.

Execution-Understanding Decoupling. Model scale substantially improves execution ($a_1 = 0.20$-0.23, $p < 0.001$ across conditions) but shows limited improvement in understanding ($a_2 = 0.04$-0.05, $p \leq 0.05$), with an efficiency gap of approximately 4:1. Under anonymous conditions, only understanding contributes significantly to reasoning ($b_2 = 1.31$, $p < 0.001$), while execution shows no significant effect ($b_1 \approx 0$, $p = 0.507$).

5.5 Cross-Task Generalization

Experimental Setup. To investigate the extent to which II capability influences downstream reasoning performance, we compute correlations between II performance and a diverse set of downstream reasoning benchmarks, including ProofWriter (Tafjord

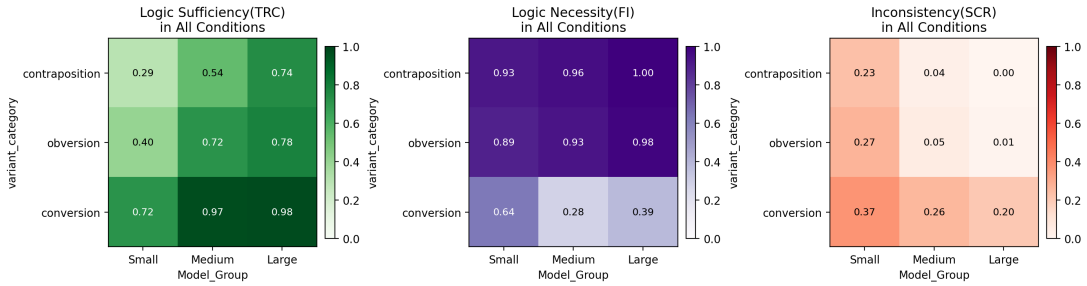


Figure 4: Conditional probability metrics. **Large models achieve high causal sufficiency while small models rely more heavily on semantic compensation.**

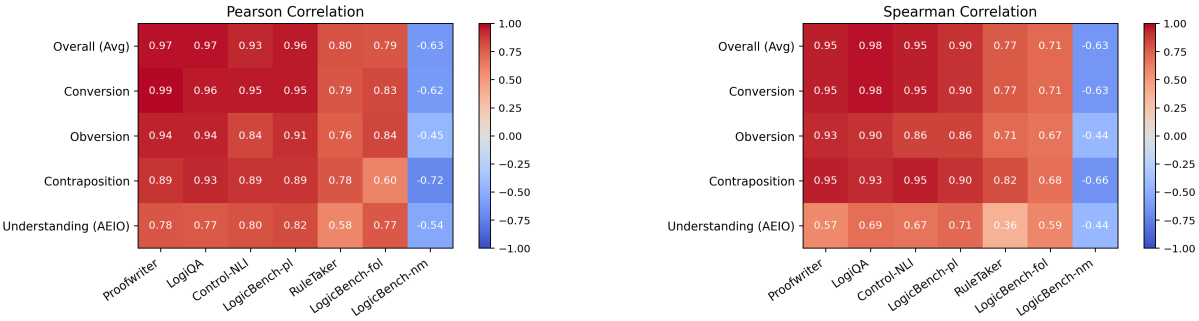


Figure 5: Correlation heatmaps between immediate inference capabilities and downstream reasoning benchmarks.

et al., 2021), LogiQA (Liu et al., 2021b), ConTRoL (Liu et al., 2021a), RuleTaker (Clark et al., 2020), LogicBench-PL, LogicBench-FOL, and LogicBench-NM (Parmar et al., 2024). Figure 5 presents the Pearson and Spearman correlation heatmaps. More results and detailed analyses are provided in Appendix F.1.

Strong positive correlations support the foundational role of II. II demonstrates strong and consistent positive associations with a broad range of formal reasoning benchmarks. In particular, the overall II score tracks closely with downstream performance on ConTRoL ($r = 0.93$, $\rho = 0.95$), ProofWriter ($r = 0.97$, $\rho = 0.95$), and LogiQA ($r = 0.97$, $\rho = 0.98$), and the transformation operations and opposition relations show the same general pattern. These results suggest that complex multi-step reasoning is not an independently emergent capability, but instead rests on more elementary II abilities: models that can reliably execute II also tend to perform better on downstream formal reasoning tasks.

Negative correlations reveal a trade-off with non-monotonic reasoning. In contrast to the positive correlations observed for formal reasoning benchmarks, II shows a consistent negative relationship with non-monotonic reasoning. The overall II metric is negatively correlated with LogicBench-NM ($r = -0.63$, $\rho = -0.63$), and the same pattern holds across all transformation operations and opposition relations. This result reveals that the two reasoning paradigms rely on distinct underlying capabilities. The traditional logic is monotonic, meaning that adding new premises cannot invalidate previously established conclusions. Non-monotonic reasoning, by contrast, requires models to retract prior conclusions upon receiving new information. Models that excel at the former may find their tendency toward rigid rule execution interfering with the flexible reasoning mechanisms demanded by the latter.

6 Model Training & Experiments

We design controlled training intervention experiments on two base models: Qwen-2.5-7B and Llama-3.1-8B. To isolate the effect of different types of foundational capability training, we use only II, syllogism, or first-order logic (FOL) data for training, without incorporating any other reasoning data. We compare the following conditions: (1) Baseline: no training; (2) Syllogism-Only: training exclusively with syllogistic examples; (3) II-Only: training exclusively with II examples; (4) FOL-Only: training exclusively on FOL data; (5) Mixed-Full: training with both II and syllogistic data; (6) Mixed-Reduced: training with the full

Condition	II *Acc*	Syllogism *Acc*	ProofWriter *Deduction*	LogiQA *MCQ*	ConTRoL *NLI*	RuleTaker *Rule-based*	LogicBench *(PL)*	LogicBench *(FOL)*	StrategyQA *Acc*
Baseline	39.27	35.00	57.67	51.67	56.00	64.00	69.67	69.67	66.00
Syllogism-Only	33.78 (-5.49)	77.08 (+42.08)	63.00 (+5.33)	51.67 (+0.00)	55.00 (-1.00)	63.00 (-1.00)	68.67 (-1.00)	73.00 (+3.33)	68.00 (+2.00)
Mixed-Reduced	76.84 (+37.57)	78.96 (+43.96)	61.00 (+3.33)	52.67 (+1.00)	56.67 (+0.67)	61.00 (-3.00)	68.00 (-1.67)	71.33 (+1.66)	**69.33 (+3.33)**
Mixed-Full	78.92 (+39.65)	**79.79 (+44.79)**	61.33 (+3.66)	52.00 (+0.33)	54.33 (-1.67)	61.33 (-2.67)	69.00 (-0.67)	73.33 (+3.66)	68.00 (+2.00)
FOL-Only	44.94 (+5.67)	42.29 (+7.29)	**65.00 (+7.33)**	51.33 (-0.33)	**57.33 (+1.33)**	64.33 (+0.33)	**83.67 (+14.00)**	**94.00 (+24.33)**	66.67 (+0.67)
II-Only	**91.35 (+52.08)**	32.08 (-2.92)	58.67 (+1.00)	**53.00 (+1.33)**	56.33 (+0.33)	**67.33 (+3.33)**	70.00 (+0.33)	73.33 (+3.66)	67.33 (+1.33)

(a) Qwen-2.5-7B

Condition	II *Acc*	Syllogism *Acc*	ProofWriter *Deduction*	LogiQA *MCQ*	ConTRoL *NLI*	RuleTaker *Rule-based*	LogicBench *(PL)*	LogicBench *(FOL)*	StrategyQA *Acc*
Baseline	31.27	33.54	53.67	46.00	46.00	60.67	62.67	68.67	60.33
Syllogism-Only	35.00 (+3.73)	**80.83 (+47.29)**	66.33 (+12.66)	46.00 (+0.00)	43.33 (-2.67)	**65.00 (+4.33)**	68.67 (+6.00)	79.00 (+10.33)	**67.00 (+6.67)**
Mixed-Reduced	73.94 (+42.67)	76.88 (+43.34)	**69.33 (+15.66)**	**47.33 (+1.33)**	44.00 (-2.00)	**65.00 (+4.33)**	64.00 (+1.33)	76.33 (+7.66)	66.00 (+5.67)
Mixed-Full	79.12 (+47.85)	78.96 (+45.42)	69.00 (+15.33)	46.33 (+0.33)	46.33 (+0.33)	62.00 (+1.33)	66.67 (+4.00)	77.00 (+8.33)	63.67 (+3.34)
FOL-Only	39.31 (+8.04)	35.83 (+2.29)	56.00 (+2.33)	38.33 (-7.67)	40.00 (-6.00)	64.00 (+3.33)	**83.33 (+20.66)**	**98.00 (+29.33)**	66.00 (+5.67)
II-Only	**90.47 (+59.20)**	31.87 (-1.67)	63.67 (+10.00)	46.33 (+0.33)	**48.67 (+2.67)**	**65.00 (+4.33)**	63.33 (+0.66)	74.33 (+5.66)	64.67 (+4.34)

(b) Llama-3.1-8B

Table 6: Training intervention results across different reasoning tasks. Best results are **bolded**, second-best is underlined, red indicates negative transfer. **II-Only training provides the most robust cross-task transfer.**

II corpus and 50% of the syllogistic data used in Mixed-Full.

We evaluate on seven downstream reasoning benchmarks: ProofWriter, LogiQA, ConTRoL, RuleTaker, LogicBench-PL, LogicBench-FOL, and StrategyQA (Geva et al., 2021). The first six are formal reasoning benchmarks identified as strongly correlated with II capability in Section 5.5, while StrategyQA serves as a natural language reasoning task beyond formal logic. All evaluations are conducted under zero-shot settings. Details are provided in Appendix F, results are shown in Table 6.

II-Only training provides the most consistent positive cross-task transfer. II-Only training is the only condition that achieves positive transfer on all seven downstream benchmarks across both base models, without incurring any negative transfer. On Qwen-2.5-7B, II-Only improves ProofWriter (+1.00%), LogiQA (+1.33%), ConTRoL (+0.33%), RuleTaker (+3.33%), LogicBench-PL (+0.33%), LogicBench-FOL (+3.66%), and StrategyQA (+1.33%), with the same pattern observed on Llama-3.1-8B. Although II-Only does not always achieve the highest score on every benchmark, its consistent stability across architectures and tasks demonstrates that training on II alone can effectively improve downstream reasoning without task-specific supervision, further supporting that **II serves as a foundational capability**.

Syllogism-Only and FOL-Only training improve selected aligned tasks, but both can induce negative transfer. In contrast to II-Only, the more task-aligned supervision settings yield stronger gains on a narrower subset of benchmarks while remaining less stable overall. On Qwen-2.5-7B, Syllogism-Only improves ProofWriter (+5.33%), LogicBench-FOL (+3.33%), and StrategyQA (+2.00%), but incurs negative transfer on ConTRoL, RuleTaker, and LogicBench-PL (all -1.00%); on Llama-3.1-8B, it also harms ConTRoL (-2.67%) despite larger gains on aligned benchmarks. Likewise, FOL-Only achieves the strongest gains on LogicBench-PL and LogicBench-FOL, but still introduces negative transfer on broader tasks such as LogiQA (-7.67%) and ConTRoL (-6.00%) on Llama-3.1-8B. Taken together, these results show that task-aligned training can improve selected downstream tasks, but often at the cost of degrading broader reasoning capabilities.

7 Conclusion

In this paper, we first propose that II serves as a foundational capability for logical reasoning in LLMs. We construct IIBench, a systematic benchmark for evaluating II, and reveal that all SoTA LLMs exhibit deficiencies in this fundamental capability. Through causal and correlation analyses, we confirm the foundational role of II for downstream reasoning, while uncovering that current models struggle to flexibly switch between structural reasoning and surface pattern matching. Training experiments further validate that enhancing II effectively improves downstream reasoning performance.

Limitations

Language Coverage Our benchmark and empirical analyses are limited to English, primarily because the downstream reasoning benchmarks used for validation, including LogicBench, ProofWriter, and RuleTaker, are predominantly English-based. However, logical reasoning is not inherently language-specific, and our benchmark construction pipeline is largely language-agnostic, relying on formal logical templates and structured entity substitution rather than English-specific lexical knowledge. We do not evaluate immediate inference in typologically diverse languages or test cross-lingual transfer, so the extent to which our findings generalize across languages remains unclear.

Scope This work focuses on traditional logic, a fragment of FOL centered on standard categorical proposition forms. We do not independently evaluate propositional connectives, nested quantifiers, polyadic predicates, or non-classical logics such as modal and temporal logic. Our cross-task correlation analysis also covers only a limited set of downstream benchmarks rather than the full space of formal reasoning tasks. This restricted scope allows us to isolate foundational capabilities and establish clearer causal links.

Model Coverage We evaluate 8 representative models spanning parameter scales from 7B to frontier level, including both closed-source and open-source architectures. Although experimental results reveal consistent failure patterns across different models, suggesting that the issues are general rather than model-specific, we cannot guarantee that these findings apply to all existing or future model architectures, particularly models specifically optimized for logical reasoning. We publicly release our benchmark and evaluation code to facilitate replication and extension on new models.

Causal Analysis This work employs structural equation modeling for causal mediation analysis, which relies on specific causal graph assumptions. Causal relationships are established based on observational data rather than randomized controlled experiments, and alternative causal structures (e.g., a common latent capability simultaneously affecting both immediate inference and syllogistic reasoning) cannot be entirely ruled out. To enhance the reliability of our conclusions, we provide mutual validation through six experiments examining overall performance, robustness, failure mechanisms, causal effects, and cross-task generalization. Causal invariance tests further verify that large models exhibit consistent causal structures across different semantic conditions.

Training Validation Due to computational constraints, our training intervention experiments are limited to smaller models under 10B parameters, specifically Qwen-2.5-7B and Llama-3.1-8B, and do not extend to larger models. Still, the intervention results are consistent across these two architectures and align with our observational analyses on larger-scale closed-source models. Therefore, our causal conclusions regarding the foundational role of II are relatively reliable. Whether these exact transfer dynamics persist in larger models remains an open question.

Ethics Statement

Data Source and Privacy All data in this research is sourced from the public knowledge graph WikiData (CC0 license) and WordNet (Princeton WordNet license). The dataset does not involve personal privacy information, user data, or copyrighted content. The benchmark dataset and code will be publicly released to promote reproducibility.

Factuality, Toxicity, and Bias We implement multiple safeguards during data construction to improve quality and safety. Factual instances are grounded in community-verified entity-class relations in Wikidata, while counterfactual and anonymous instances are designed to test logical reasoning independently of world knowledge. All entity-class pairs are manually reviewed to exclude offensive, controversial, or sensitive content. Labels are generated by logical rules rather than human annotation, reducing annotator bias. Although Wikidata and WordNet may contain coverage biases, logical validity in our benchmark is determined by formal rules rather than semantic preference.

AI-Assisted Writing This paper used Claude-Sonnet 4.5 for language polishing. All research design, experimentation, analysis, and core arguments were completed by the authors. AI assistance was limited to improving the fluency and academic style and did not involve substantive contributions to the research content.

References

David Paul Ausubel, Joseph Donald Novak, Helen Hanesian, and 1 others. 1978. Educational psychology: A cognitive view.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational linguistics*, 39(3):463–472.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Irving M Copi, Carl Cohen, and Kenneth McMahon. 2016. *Introduction to logic*. Routledge.

Craig DeLancey. 2017. *A concise introduction to logic*. Open SUNY Textbooks.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. 2024. Logical consistency of large language models in fact-checking. *CoRR*.

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2024. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 14852–14882.

Laurence R Horn. 2006. Contradiction.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7811–7818.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, and 1 others. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

William Calvert Kneale and Martha Kneale. 1984. *The development of logic*. Oxford University Press.

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

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021a. Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396.

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. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pages 221–242. Springer.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4885–4901.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. Linc: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176.

Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. Logicinference: A new dataset for teaching logical inference to seq2seq models. *CoRR*, abs/2203.15099.

OpenAI. 2024. Learning to reason with LLMs. *OpenAI Blog*.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada.

2024. Exploring reasoning biases in large language models through syllogism: Insights from the neubaroco dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16063–16077.

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

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707.

Terence Parsons. 2014. *Articulating medieval logic*. OUP Oxford.

Terence Parsons and Graziana Ciola. 1997. The traditional square of opposition.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 856–865.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8713–8721.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3762–3780.

Robert S Siegler. 1998. *Emerging minds: The process of change in children's thinking*. Oxford University Press.

R. Smith. 1989. *Prior Analytics*. HPC Classics Series. Hackett Publishing Company.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Alfred Tarski. 1956. The concept of truth in formalized languages.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. Logicasker: Evaluating and improving the logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pages 1112–1122.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 250–255.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The eleventh international conference on learning representations*.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. *arXiv preprint arXiv:2104.06598*.

Shi Zong and Jimmy Lin. 2024. Categorical syllogisms revisited: A review of the logical reasoning abilities of llms for analyzing categorical syllogisms. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 230–239.

A Extended Related Work

A.1 Logical Reasoning Benchmarks

The evaluation of logical reasoning in language models has evolved substantially over recent years. Early benchmarks focused on specific reasoning types: LogiQA (Liu et al., 2021b), ReClor (Yu et al., 2020), and AR-LSAT (Zhong et al., 2021) assess multi-step syllogistic and analytical reasoning drawn from standardized tests, while NeuBAROCO (Ozeki et al., 2024) specifically targets categorical syllogisms in natural language. For deductive inference, ProofWriter (Tafjord et al., 2021) evaluates multi-hop reasoning with explicit proof generation, and FOLIO (Han et al., 2024) provides human-authored problems requiring first-order logic reasoning.

Natural language inference benchmarks have progressed from SNLI's (Bowman et al., 2015) crowd-sourced sentence pairs through MNLI's (Williams et al., 2018) multi-genre coverage to ANLI's (Nie et al., 2020) adversarially constructed examples designed to challenge model weaknesses. Compositional generalization has been systematically probed through SCAN's (Lake and Baroni, 2018) command-to-action mappings, COGS's (Kim and Linzen, 2020) semantic parsing tasks, and CFQ's (Keysers et al., 2019) question-answering requiring novel predicate-argument combinations.

More recent efforts have pursued unified evaluation frameworks. LogicBench (Parmar et al., 2024) provides comprehensive coverage across propositional logic, first-order logic, and non-monotonic reasoning within a single benchmark. Several studies have specifically examined large language model performance on categorical syllogisms (Eisape et al., 2024; Zong and Lin, 2024), revealing systematic errors in quantifier interpretation and premise integration that persist even in state-of-the-art models.

Despite this extensive coverage, existing benchmarks share a critical limitation: they do not isolate immediate inference for independent evaluation. Transformation operations such as conversion, obversion, and contraposition, are assessed only implicitly as embedded components of multi-step tasks. Furthermore, benchmarks typically evaluate performance on single formulations, leaving uninvestigated whether models recognize logically equivalent reformulations. Our work addresses these gaps by constructing a benchmark that systematically evaluates immediate inference across controlled equivalent formulations, enabling precise diagnosis of foundational logical capabilities.

A.2 Reasoning Failure Analysis

Research into reasoning failures has identified multiple sources of model brittleness. Work on spurious correlations and shortcut learning (McCoy et al., 2019; Gururangan et al., 2018) has demonstrated that models frequently exploit annotation artifacts, lexical overlap, and superficial heuristics rather than engaging in genuine logical reasoning. These findings have prompted development of challenge sets specifically designed to expose such shortcuts.

Studies of semantic consistency (Ribeiro et al., 2019) have revealed that models exhibit unstable predictions when inputs are paraphrased in meaning-preserving ways, suggesting that apparent reasoning success may reflect pattern matching to specific surface forms rather than robust semantic understanding. Adversarial perturbation research (Jia and Liang, 2017; Ribeiro et al., 2018) has further demonstrated that minor modifications—inserting irrelevant sentences, altering

named entities, or introducing distractors—can dramatically degrade performance on reasoning tasks.

Investigations into quantifier understanding and monotonicity reasoning (Yanaka et al., 2019; Richardson et al., 2020; Geiger et al., 2020) have uncovered systematic difficulties in handling logical operators fundamental to natural language inference. Models struggle with quantifier scope ambiguities, fail to properly track monotonicity properties through sentence embeddings, and exhibit inconsistent behavior on inferences involving negation.

Research on logical consistency has revealed striking failures: models trained on "A is B" systematically fail to infer "B is A" (Berglund et al., 2023), indicating fundamental limitations in bidirectional reasoning. Recent studies have extended consistency probing to the propositional level (Ghosh et al., 2024) and conducted controlled experiments analyzing systematic reasoning errors in frontier models (Nezhurina et al., 2024).

While these analyses have substantially advanced understanding of model limitations, they share two characteristics that our work addresses. First, they predominantly examine semantic equivalence (e.g., "All birds fly" versus "Every bird can fly") rather than logical structural transformations where propositional form changes while truth conditions are preserved (e.g., "All S are P" versus "No S are non-P" via obversion).Second, they establish correlational patterns between observed failures and hypothesized causes, but do not provide causal evidence quantifying how specific foundational deficits propagate to downstream reasoning performance. Our causal mediation analysis directly addresses this gap, establishing immediate inference as a significant causal mediator for syllogistic reasoning, with near-perfect correlation (ρ = 0.98) across formal reasoning benchmarks.

A.3 Reasoning Improvement Methods

Methods for improving reasoning capabilities have developed along several trajectories. Prompting-based approaches began with chain-of-thought prompting (Wei et al., 2022), which elicits intermediate reasoning steps through few-shot demonstrations, and extended to zero-shot variants (Kojima et al., 2022) that trigger reasoning through simple instructions and automatic methods (Zhang et al., 2022) that generate chain-of-thought examples programmatically.

Fine-tuning approaches have targeted reasoning capabilities through various strategies. STaR (Zelikman et al., 2022) bootstraps reasoning ability by training on self-generated rationales, while knowledge distillation (Ho et al., 2023) transfers reasoning capabilities from larger to smaller models. Instruction tuning on diverse reasoning tasks (Chung et al., 2024) has shown benefits for generalization.

Neuro-symbolic methods seek to combine neural language models with symbolic reasoning engines. LINC (Olausson et al., 2023) exemplifies this approach by coupling language models with first-order logic provers, delegating formal inference to systems with guaranteed soundness.

Self-consistency methods (Wang et al., 2022) improve reliability by sampling multiple reasoning paths and selecting answers through majority voting or other aggregation mechanisms, exploiting the intuition that correct reasoning should be more reproducible than incorrect reasoning.

Recent advances have introduced several innovations. Process reward models (Lightman et al., 2023) provide fine-grained supervision at individual reasoning steps rather than only final answers, enabling more precise credit assignment during training. Math-Shepherd (Wang et al., 2024) extends this approach with automated verification that does not require human step-level annotations. Research on test-time compute scaling (Snell et al., 2024) has demonstrated that allocating additional computation during inference—through search, verification, or iterative refinement—can yield improvements competitive with or exceeding those from scaling model parameters. Reasoning-specialized models such as o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025) represent architectural and training innovations specifically targeting complex reasoning.

These diverse methods share an implicit assumption: that models possess robust foundational logical capabilities which need only be better orchestrated, elicited, or supervised. Our findings challenge this assumption. We demonstrate that immediate inference—the most elementary form of logical reasoning—exhibits systematic deficiencies even in state-of-the-art models. When these foundational capabilities are brittle across logically equivalent formulations, methods that build upon them inherit corresponding limitations. Moreover, our discovery of a fundamental trade-off between structural reasoning and surface pattern matching suggests that current approaches face inherent constraints in resolving conflicts between reasoning

modes, pointing toward architectural innovations as a direction for future work.

B Scope and Positioning

Starting Point This work takes traditional logic as its starting point, focusing on the foundational components of categorical proposition reasoning. We adopt a broad conception of "immediate inference" that encompasses two types of elementary operations:

- **Transformation Operations**: Operations that derive logically equivalent propositions from a single premise, including conversion, obversion, and contraposition. In traditional logic textbooks, these operations are typically referred to as the core content of immediate inference (Copi et al., 2016; DeLancey, 2017).
- **Opposition Relations**: Truth-value dependencies between propositions sharing identical subject and predicate terms but differing in quantity, quality, or both. These include four relations: contradictory, contrary, subcontrary, and subaltern, collectively organized in the traditional Square of Opposition (Parsons and Ciola, 1997; Horn, 2006).

While traditional textbooks typically discuss transformation operations and opposition relations separately, both involve reasoning from a single categorical proposition and together constitute the foundational capabilities for categorical proposition reasoning. Transformation operations test models' *execution capability* (whether they can correctly perform logical transformations), while opposition relations test models' *understanding capability* (whether they can correctly judge truth-value dependencies between propositions).

Positioning This work targets immediate inference as the foundational layer of categorical-proposition reasoning, jointly covering transformation operations (execution) and opposition relations (understanding). Formal definitions appear in Section 3 and complete specifications in Appendix C; the remainder of this appendix delineates what lies outside our scope.

Out of Scope This work does not address the following logical systems and reasoning types: propositional logic connectives ($\wedge, \vee, \neg, \rightarrow$), which involve different mechanisms from categorical propositions; full first-order logic involving polyadic predicates, nested quantifiers, and relational reasoning, which extends beyond traditional logic; higher-order logic, which quantifies over predicates and functions; modal logic involving necessity and possibility operators; temporal logic involving temporal relations; and mathematical reasoning, which although deductive involves numerical computation with different underlying mechanisms.

C Immediate Inference: Complete Specification

This appendix provides comprehensive details on immediate inference, including validity justifications for transformation operations, complete specifications of opposition relations, distinctions from related concepts, and natural language variations.

C.1 Transformation Operations: Validity Justification

C.1.1 Why Conversion Is Valid Only for E and I Types

Conversion exchanges the subject and predicate terms while preserving the quantifier and quality. Its validity depends on whether the original proposition expresses a symmetric relation between the subject and predicate classes (Copi et al., 2016; DeLancey, 2017).

Valid Cases (E and I Types) **E-type** (No S are P $\rightarrow$ No P are S): Universal negatives express complete mutual exclusion. If no members of S belong to P, then no members of P belong to S—the relation is perfectly symmetric. For example, "No fish are mammals" logically entails "No mammals are fish."

I-type (Some S are P $\rightarrow$ Some P are S): Particular affirmatives assert existential overlap. If at least one thing is both S and P, then at least one thing is both P and S—existential claims are inherently symmetric. For example, "Some animals are mammals" logically entails "Some mammals are animals."

Invalid Cases (A and O Types) **A-type**: Universal affirmatives express asymmetric subset containment. "All mammals are animals" means the class of mammals is contained within the class of animals, but the converse does not hold—not all animals are mammals. The subset relation is directional and cannot be reversed. Therefore, standard A$\rightarrow$A conversion has no valid form. However,

there exists a special case: *conversion by limitation* (A→I) is valid (All S are P → Some P are S), because the subset relation entails existential overlap (Parsons, 2014).

O-type: Particular negatives express partial exclusion from a specific perspective. "Some animals are not mammals" means there exist animals outside the mammal class, but this does not imply the existence of mammals outside the animal class. Therefore, O-type conversion has no valid form.

C.1.2 Why Obversion Is Universally Valid

Obversion changes the quality of a proposition (affirmative ↔ negative) while replacing the predicate with its complement. This transformation exploits the fundamental logical equivalence: $x \in P \Leftrightarrow x \notin \overline{P}$, where $\overline{P}$ denotes the complement of P (Copi et al., 2016; Kneale and Kneale, 1984).

This equivalence holds regardless of quantifier scope:

- **A-type**: All S are P → No S are non-P (if every S belongs to P, then no S belongs to non-P)
- **E-type**: No S are P → All S are non-P (if no S belongs to P, then every S belongs to non-P)
- **I-type**: Some S are P → Some S are not non-P (if at least one S belongs to P, then at least one S does not belong to non-P)
- **O-type**: Some S are not P → Some S are non-P (if at least one S does not belong to P, then at least one S belongs to non-P)

Because obversion is grounded in the definitional relationship between a class and its complement—a relationship that is logically necessary and independent of empirical content—it remains valid across all proposition types.

C.1.3 Why Contraposition Is Valid Only for A and O Types

Contraposition can be understood as the composition of obversion followed by conversion (Copi et al., 2016). Given a proposition, we first apply obversion (which is always valid), then apply conversion to the result. The validity of contraposition therefore depends on whether conversion is valid for the obverted form.

Valid Cases (A and O Types) **A-type** (All S are P → All non-P are non-S): Obversion yields "No S are non-P" (E-type), and conversion is valid for E-type propositions, yielding "No non-P are S." Further obversion gives "All non-P are non-S." This works because A-type propositions express universal implication: if S implies P, then not-P implies not-S (the contrapositive law) (Kneale and Kneale, 1984).

O-type (Some S are not P → Some non-P are not non-S): Obversion yields "Some S are non-P" (I-type), and conversion is valid for I-type propositions, yielding "Some non-P are S." Further obversion gives "Some non-P are not non-S."

Invalid Cases (E and I Types) **E-type**: Obversion yields "All S are non-P" (A-type), but conversion is invalid for A-type propositions. Therefore, E-type contraposition has no valid form.

I-type: Obversion yields "Some S are not non-P" (O-type), but conversion is invalid for O-type propositions. Therefore, I-type contraposition has no valid form.

The validity pattern reflects a deep duality in propositional logic: A and O types are duals under negation, and both preserve their logical structure under the contrapositive transformation (Parsons, 2014).

C.1.4 Complete Transformation Specification

Table 7 provides the complete specification of all transformation operations, including logical forms, concrete examples, and validity status.

C.2 Opposition Relations: Complete Specification

C.2.1 Structure of the Square of Opposition

The Square of Opposition defines truth-value dependencies between propositions that share the same subject and predicate terms but differ in quantity (universal/particular) or quality (affirmative/negative) (Parsons and Ciola, 1997; Horn, 2006).

Four fundamental relations constitute the square:

Contradictory Relates propositions differing in both quantity and quality (A-O and E-I pairs). Contradictory propositions cannot both be true and cannot both be false; the truth of one necessarily implies the falsity of the other.

Contrary Relates universal propositions of opposite quality (A-E pairs). Contrary propositions

Type	Original Form	Operation	Result Form	Example	Valid
A	All S are P	Conversion	—	—	✗
		Conversion (limited)	Some P are S (I)	Some animals are mammals	✓
		Obversion	No S are non-P (E)	No mammals are non-animals	✓
		Contraposition	All non-P are non-S (A)	All non-animals are non-mammals	✓
E	No S are P	Conversion	No P are S (E)	No mammals are fish	✓
		Obversion	All S are non-P (A)	All fish are non-mammals	✓
		Contraposition	—	—	✗
I	Some S are P	Conversion	Some P are S (I)	Some mammals are animals	✓
		Obversion	Some S are not non-P (O)	Some animals are not non-mammals	✓
		Contraposition	—	—	✗
O	Some S are not P	Conversion	—	—	✗
		Obversion	Some S are non-P (I)	Some animals are non-mammals	✓
		Contraposition	Some non-P are not non-S (O)	Some non-mammals are not non-animals	✓

Table 7: Complete specification of transformation operations with logical forms and examples. Invalid transformations have no valid output form.

cannot both be true, though both may be false.

Subcontrary Relates particular propositions of opposite quality (I-O pairs). Subcontrary propositions cannot both be false, though both may be true.

Subaltern Relates propositions of the same quality but different quantity (A-I and E-O pairs). The truth of the universal proposition (superaltern) entails the truth of the particular proposition (subaltern), but not vice versa.

C.2.2 Logical Forms and Truth Rules

Table 8 presents the complete specification of opposition relations, including logical forms, truth rules, and illustrative examples.

C.3 Distinctions from Related Concepts

C.3.1 Immediate Inference vs. Logical Entailment

Entailment is a general directional relation where one proposition logically follows from another. Immediate inference operations establish specific logical relationships between propositions, which may be equivalences or unidirectional entailments depending on the operation and proposition type (Copi et al., 2016). Obversion produces logical equivalences for all proposition types. Conversion produces equivalences only for E and I types; for A-type propositions, only conversion by limitation (A→I) is valid, which yields unidirectional entailment rather than equivalence. Contraposition produces equivalences for A and O types. The validity patterns are summarized in Table 7.

C.3.2 Immediate Inference vs. Semantic Paraphrase

Paraphrases preserve meaning through lexical or syntactic variation while maintaining the same logical structure (Bhagat and Hovy, 2013). For example, "All birds fly" and "Every bird can fly" are paraphrases with identical logical form (universal affirmative). Immediate inference, by contrast, *transforms the logical structure itself*—changing quantifier type, quality, or term positions—while preserving truth value. "All S are P" (A, universal affirmative) and "No S are non-P" (E, universal negative) are logically equivalent despite having different structural forms.

C.3.3 Immediate Inference vs. Syllogism

Syllogistic reasoning derives a conclusion from *two premises* sharing a common middle term (e.g., "All A are B; All B are C; therefore, All A are C") (Smith, 1989). Immediate inference operates on a *single premise*, reformulating it without introducing new information. However, immediate inference often serves as a preprocessing step in syllogistic reasoning—transforming premises to match valid syllogistic figures (Copi et al., 2016).

C.3.4 Immediate Inference vs. Logical Equivalence

Logical equivalence is a property (two propositions have the same truth value in all models), while immediate inference is a *mechanism* for establishing equivalence through specific structural transformations (Tarski, 1956). Not all logical equivalences are immediate inferences—for instance, double negation elimination ("It is not the case that S are not P" ≡ "S are P") is an equivalence but not a categorical immediate inference operation.

Relation	Pair	Logical Form	Truth Rule	Example
Contradictory	A-O	All S are P ↔ Some S are not P	Exactly one true	"All dogs are mammals" (T) ⇒ "Some dogs are not mammals" (F)
	E-I	No S are P ↔ Some S are P	Exactly one true	"No birds are mammals" (T) ⇒ "Some birds are mammals" (F)
Contrary	A-E	All S are P ↔ No S are P	Cannot both be true; may both be false	"All animals are mammals" (F) ⇒ "No animals are mammals" (undetermined)
Subcontrary	I-O	Some S are P ↔ Some S are not P	Cannot both be false; may both be true	"Some animals are mammals" (T) ⇒ "Some animals are not mammals" (undetermined)
Subaltern	A→I	All S are P → Some S are P	Universal true ⇒ particular true	"All dogs are mammals" (T) ⇒ "Some dogs are mammals" (T)
	E→O	No S are P → Some S are not P	Universal true ⇒ particular true	"No fish are mammals" (T) ⇒ "Some fish are not mammals" (T)

Table 8: Complete specification of opposition relations with logical forms, truth rules, and examples.

D Data Construction Details

D.1 Seed Selection Criteria

The core challenge in immediate inference is to precisely control the subject term S and predicate term P. We construct *factual* (S, P) seed pairs from Wikidata by extracting two canonical relation types: *subclass-of* (P279) for taxonomic inclusion and *disjoint-with* (P2738) for mutual exclusivity. Concretely, we sample 200 *subclass-of* pairs and split this pool to instantiate the affirmative forms (A and I), and sample 200 *disjoint-with* pairs and split this pool to instantiate the negative forms (E and O).

Wikidata seed example (P279)


```
{
  "source": "wikidata",
  "pair": {
    "X": { "qid": "Q7569", "label": "child" },
    "Y": { "qid": "Q5", "label": "human" }
  },
  "relation": {
    "property": "P279",
    "name": "subclass of",
    "path": "wdt:P279",
    "hop": 1
  },
  "intended_forms": ["A", "I"],
  "evidence": "wdt:P279 (1-hop)"
}
```


Example benchmark instances with the same Wikidata seed


```
{"id": "CONV_000021", "family": "AEIO", "
    subtask": "conversion_generation", "
    premise": "All child are human.", "
    premise_form": "A", "subject_term": "child
    ", "predicate_term": "human", "
    subject_neg_depth": 0, "
    predicate_neg_depth": 0, "
    subject_canonical": "child", "
    predicate_canonical": "human", "candidate
    ": "Some human are child.", "
    candidate_form": "I", "
    candidate_subject_term": "human", "
    candidate_predicate_term": "child", "
    candidate_subject_neg_depth": 0, "
    candidate_predicate_neg_depth": 0, "
    category": "fact"}
```


To disentangle semantic influence from world knowledge, we generate *counterfactual* pairs by sampling domain-specific synonym sets from WordNet (Miller, 1995) and randomizing (S, P) pairings across domains, producing propositions that remain semantically coherent while contradicting factual knowledge. We additionally extract 100 WordNet term pairs from five controlled domains: *math* (`noun.quantity`, `noun.shape`), *relations* (`noun.relation`), *motives* (`noun.motive`), *language* (`noun.communication`), and *tops* (`noun.Tops`).

WordNet domain-check cases


```
{
  "cases": [
    {
      "term": "octillion",
      "best_target_group": "math",
      "wnid": "octillion.n.01",
      "lexname": "noun.quantity",
      "definition": "the number that is
    represented as a one followed by 27 zeros"
    },
    {
      "term": "registered_mail",
      "best_target_group": "language",
      "wnid": "registered_mail.n.01",
      "lexname": "noun.communication",
      "definition": "mail that is registered
    by the post office when sent in order to
    assure safe delivery"
    }
  ]
}
```


Example benchmark instances with the same

WordNet seed


```
{"id": "OBV_000144", "family": "AEIO", "
    subtask": "obversion_generation", "premise
    ": "Some octillion are not registered_mail
    .", "premise_form": "O", "subject_term": "
    octillion", "predicate_term": "
    registered_mail", "subject_neg_depth": 0,
    "predicate_neg_depth": 0, "
    subject_canonical": "octillion", "
    predicate_canonical": "registered_mail", "
    candidate": "Some octillion are non-
    registered_mail.", "candidate_form": "I",
    "candidate_subject_neg_depth": 0, "
    candidate_predicate_neg_depth": 1, "
    category": "masked_counterfactual"}
```


Meanwhile, we construct the *anonymous* condition by replacing S and P with randomly generated lowercase nonce strings, thereby removing lexical-semantic content.

For the syllogistic reasoning portion of the benchmark, we instantiate 24 valid syllogism forms spanning the four classical figures. For each form, we sample two factual (S, M, P) realizations, yielding 48 FACTUAL triplets in total. In addition, we construct 24 ANONYMOUS and 24 COUNTERFACTUAL triplets using the same form set, resulting in 96 (S, M, P) triplets per diagnostic category. All triplets are assembled exclusively from the same Wikidata *subclass-of* and *disjoint-with* pools, ensuring that both premises are grounded in verified taxonomic or disjointness constraints. For any form requiring existential import (i.e., involving I or O premises), we additionally enforce a non-emptiness constraint by querying Wikidata for a witness instance via WDQS, rejecting candidates without evidence of existence.

Syllogistic seed example


```
{
  "form_id": "AAI-1",
  "figure": 1,
  "category": "factual",
  "triplet": {
    "S": { "qid": "Q25381237", "label": "Rugby
     $\grave{a}$ 5", "label_lang": "en" },
    "M": { "qid": "Q5849", "label": "rugby
    union", "label_lang": "en" },
    "P": { "qid": "Q5378", "label": "rugby", "
    label_lang": "en" }
  },
  "construction": {
    "form_id": "AAI-1",
    "figure": 1,
    "major_premise": { "quantifier": "A", "
    subject_qid": "Q5849", "predicate_qid": "
    Q5378" },
    "minor_premise": { "quantifier": "A", "
    subject_qid": "Q25381237", "predicate_qid
    ": "Q5849" }
  }
}
```


Example benchmark instances with the same Wikidata seed


```
{"family": "syllogism", "subtask": "generation
    ", "split": "test", "premise1": "All rugby
     union are rugby.", "premise1_form": "A",
    "premise1_subject_term": "rugby union", "
    premise1_predicate_term": "rugby", "
    premise1_subject_neg_depth": 0, "
    premise1_predicate_neg_depth": 0, "
    premise1_subject_canonical": "rugby union
    ", "premise1_predicate_canonical": "rugby
    ", "premise2": "All Rugby $\grave{a}$ 5
    are rugby union.", "premise2_form": "A", "
    premise2_subject_term": "Rugby $\grave{a}$
     5", "premise2_predicate_term": "rugby
    union", "premise2_subject_neg_depth": 0, "
    premise2_predicate_neg_depth": 0, "
    premise2_subject_canonical": "Rugby $\
    grave{a}$ 5", "
    premise2_predicate_canonical": "rugby
    union", "candidate_gold": "Some Rugby $\
    grave{a}$ 5 are rugby.", "candidate_form":
     "I", "candidate_subject_term": "Rugby $\
    grave{a}$ 5", "candidate_predicate_term":
    "rugby", "candidate_subject_neg_depth": 0,
     "candidate_predicate_neg_depth": 0, "
    candidate_subject_canonical": "Rugby $\
    grave{a}$ 5", "
    candidate_predicate_canonical": "rugby", "
    mood": "AAI-1", "figure": 1, "
    source_instance": {"form_id": "AAI-1", "
    figure": 1, "S_qid": "Q25381237", "M_qid":
     "Q5849", "P_qid": "Q5378", "S_label": "
    Rugby $\grave{a}$ 5", "M_label": "rugby
    union", "P_label": "rugby", "construction
    ": {"form_id": "AAI-1", "figure": 1, "
    major": ["A", "Q5849", "Q5378"], "minor":
    ["A", "Q25381237", "Q5849"]}, "
    S_label_lang": "en", "M_label_lang": "en",
     "P_label_lang": "en"}, "id": "SYL_000041
    ", "base_id": "SYL_000041", "category": "
    standard", "condition": "fact", "
    transformation_note": "identity"}
```


Quality Filters. We strictly adhere to Wikidata relation semantics (P279 for inclusion and P2738 for exclusion), and manually remove pairs that are meaningless, ill-formed, or clearly incorrect. We further filter surface realizations by discarding missing or uninformative labels (e.g., raw QIDs) and labels containing temporal expressions (e.g., weekday/month patterns). For syllogism triplets, we additionally require S, M, and P to be pairwise distinct and remove duplicate instantiations.

D.2 Sample Generation Process

Prompt Format Variations. For each immediate-inference generation case (obversion, conversion, contraposition explains), we

evaluate instruction-following robustness using three controlled prompt formats that differ only in the amount of scaffolding while preserving identical output constraints (i.e., a single categorical sentence in a fixed template, or a fixed "No valid {*Operation*}" response when the operation is invalid):

- **Format 1: Direct instruction with strict output templates**

```
``You are given one categorical premise
    in classical categorical logic.
    Apply the immediate inference
    operation: Conversion.

    If a valid conversion exists in
    classical categorical logic, output
    exactly ONE
    sentence in ONE of the following four
    templates (end with a period):
    1) All <term> are <term>.
    2) No <term> are <term>.
    3) Some <term> are <term>.
    4) Some <term> are not <term>.

    If no valid conversion exists, output
    exactly:
    No valid conversion.

    Do not output any additional text.

    Premise: <premise>''
```

- **Format 2: Definition-augmented instruction**

```
``Definition: Conversion exchanges the
    subject and predicate.
    In classical categorical logic: E and
    I convert simply; A converts per
    accidens;
    O has no valid conversion.

    Apply the immediate inference
    operation: Conversion.

    [Same strict output templates /
    invalid-operation string as Format
    1.]''
```

- **Format 3: Definition + single illustrative example**

```
``Definition: Conversion exchanges the
    subject and predicate.
    Example: `All cats are mammals.' -> `
    Some mammals are cats.'

    Now apply Conversion to the given
    premise.

    [Same strict output templates /
    invalid-operation string as Format
    1.]''
```

We apply these variations only to the immediate-inference generation tasks. For completeness, AEIO truth-judgment and syllogism generation each use a single fixed prompt format:

- **AEIO Truth-Judgment (single fixed format)**

```
``Assume the premise below is true in
    classical categorical logic.
    Determine the truth status of the
    candidate statement relative to the
    premise.

    Output exactly ONE of the following
    labels (case-sensitive), and nothing
    else:
    True (the candidate must be true given
     the premise),
    False (the candidate must be false
    given the premise),
    Undetermined (otherwise).

    Do not output any additional text.

    Premise: <premise>.
    Candidate: <candidate>.''
```

- **Syllogism Generation (single fixed format)**

```
``You are given two categorical premises
    in classical syllogistic form.
    Derive the categorical relationship
    that is logically entailed by these
    premises.

    Output exactly ONE sentence in ONE of
    the following four templates (end
    with a period):
    (1) All <term> are <term>.
    (2) No <term> are <term>.
    (3) Some <term> are <term>.
    (4) Some <term> are not <term>.

    Do not output any additional text.

    Premise 1: <premise1>.
    Premise 2: <premise2>.''
```

D.3 Quality Control

All model outputs are validated and scored by an automated evaluation runner that logs both raw and normalized responses and applies task-specific scoring rules. This scripted pipeline serves as our

primary quality control mechanism for ensuring consistent, reproducible evaluation across models and prompt formats.

Stage 1: Output normalization and template compliance. Before scoring, we apply light normalization to remove formatting noise while preserving logical content: we strip common leading prefixes (e.g., "Answer:"), normalize whitespace and trailing punctuation, and standardize quotation marks. For categorical statements, we canonicalize term-level negation by treating `not` X and `non-`X as equivalent and cancelling double negation (e.g., `not non-`$X \equiv X$). Outputs are expected to conform to one of the four categorical templates (All/No/Some/Some-not) or the task-specific invalid-operation string (e.g., "No valid conversion."). Responses that do not parse into a well-formed categorical statement are marked incorrect unless they normalize to the gold target under the rules below.

Stage 2: Task-specific scoring rules. We adopt different scoring criteria depending on the task, matching the evaluation semantics implemented in our scripts.

- **AEIO truth-judgment (exact-match labels).** We score by exact match after label normalization to one of {`True`, `False`, `Undetermined`}. Common variants (e.g., `T/F/U`, `unknown`, `not sure`) are mapped to the canonical label; any other output is treated as incorrect.
- **Immediate-inference generation (canonical-form match).** For obversion, conversion, and contraposition, we score by matching the model output to the gold target after canonicalizing the categorical form into a structured representation `<quantifier>|<subject>|<predicate>` with negation normalized as above. This comparison tolerates minor surface differences (case, spacing, punctuation, and `not` vs. `non-`) but requires the same quantifier and the same normalized subject/predicate content. For invalid operations, the gold is a fixed string (e.g., "No valid conversion."), and only that normalized form is accepted.
- **Syllogism generation (equivalence closure).** For syllogism, we do *not* require the model to match the gold conclusion in a single surface form. Instead, we accept any conclusion that is equivalent to the gold under an equivalence closure of truth-preserving immediate inferences (obversion, valid conversion, valid contraposition) and their compositions. Concretely, we generate a bounded closure set of canonical forms from the gold conclusion and count the model output as correct if its canonical form appears in this set.

Stage 3: Reproducible logging and auditability. For each task and model, the runner writes per-sample JSONL records containing the item ID, gold answer, raw response, normalized response, correctness flag, prompt format identifier, and all dataset metadata fields. It also produces per-task summary files and uses state checkpoints to support interruption-safe resumption. This design enables systematic error analysis and post-hoc auditing of any evaluated instance.

D.4 Data Statistics

The complete benchmark, including all 4,900 core instances, and 384 downstream diagnostic instances constructed from valid syllogistic seeds under the Syllogistic Transformation Task (Section 4.2). We provide data in JSON format with standardized fields for easy integration with existing evaluation frameworks.

E Experimental Details

E.1 Experiment 1: Overall Performance and Systematic Deficiencies

E.1.1 Experimental Setup

We evaluate 8 representative LLMs spanning both closed-source and open-source models across different parameter scales to ensure comprehensive coverage of the current model landscape.

All experiments are conducted under zero-shot conditions using factual data to assess models' intrinsic logical reasoning capabilities without prompting interventions or semantic confounds. Sample sizes for evaluation are shown in Table 9 and Table 10.

Evaluation follows the automated, three-stage quality-control protocol in Appendix D.3.

E.1.2 Finding 1: No Model Achieves Reliability Standards

Logical reasoning demands certainty, which means a system that correctly applies logical rules 95%

Number	Conversion				Obversion				Contraposition				By Premise				Overall
	A→I	E→E	I→I	O→Inv	A→E	E→A	I→O	O→I	A→A	E→Inv	I→Inv	O→O	A	E	I	O	
#	**408**	**396**	**396**	**100**	**300**	**300**	**300**	**300**	**600**	**50**	**50**	**600**	**1308**	**746**	**746**	**1000**	**3800**

Table 9: Sample sizes for transformation operation evaluation.

Number	Premise A			Premise E			Premise I			Premise O			Overall
	A→E *Det.*	A→I *Det.*	A→O *Det.*	E→A *Det.*	E→I *Det.*	E→O *Det.*	I→A *Und.*	I→E *Det.*	I→O *Und.*	O→A *Det.*	O→E *Und.*	O→I *Und.*	
#	**101**	**100**	**94**	**87**	**89**	**77**	**101**	**98**	**84**	**89**	**85**	**95**	**1100**

Table 10: Sample sizes for opposition relation evaluation.

of the time is not "95% logical" but fundamentally unreliable. Unlike tasks where partial success indicates partial capability, logical inference requires consistent application of rules across all valid instances. Our evaluation reveals that even the strongest models fail to meet this reliability standard.

Closed-Source Models Show Near-Ceiling Performance with Catastrophic Exceptions. Gemini-2.5-Pro achieves the highest overall performance on transformation operations (98.4%) and opposition relations (99.8%), approaching but not reaching perfect reliability. Even this top-performing model exhibits a catastrophic failure: accuracy collapses to 10.0% on contraposition E→Inv. This represents a 89.8% gap from its average performance, indicating that near-perfect overall accuracy can mask complete failure on specific logical operations.

GPT-o3 demonstrates a similar pattern with 97.9% on transformation operations but a substantial drop to 77.1% on opposition relations. Within opposition relations, GPT-o3 achieves only 22.8% on the determinate A→E contrary relation and 35.6% on E→A, despite these being logically straightforward inferences. GPT-4.1 shows even more pronounced failures, achieving only 2.0% on A→I conversion and 6.0% on contraposition E→Inv.

Open-Source Models Exhibit Pervasive Deficiencies. Open-source models perform substantially lower across the board, with transformation operation accuracy ranging from 16.4% (Llama-3.1-8B) to 43.8% (Qwen-2.5-72B).

These results demonstrate that immediate inference deficiencies are pervasive across model architectures, scales, and training paradigms. No evaluated model achieves the consistency required for reliable logical reasoning.

E.1.3 Finding 2: Models Apply Unstable Logical Standards

A robust logical reasoner should exhibit consistent performance across tasks of comparable complexity. If a model understands the logical principle underlying an operation, it should apply that principle uniformly. However, our evaluation reveals systematic *task difficulty inversion*: models frequently perform worse on logically simpler tasks than on more complex ones, suggesting fundamental instability in their reasoning processes.

Opposition Relations: Undetermined Cases Outperform Determinate Cases. In the square of opposition, determinate cases require straightforward rule application, while undetermined cases require recognizing the limits of valid inference. Counterintuitively, GPT-o3 achieves near-perfect accuracy (98.8%-100.0%) on undetermined inference types yet only 22.8%-38.0% on determinate types. This 60%-77% gap suggests that GPT-o3 defaults to uncertainty responses rather than executing logical derivations when definitive judgments are required.

GPT-4.1 exhibits an even more striking inversion: it achieves 97.9% on the A→O contradictory relation but only 16.0% on the A→I subalternation. The A→I inference is logically simpler, yet GPT-4.1 fails on it 84.0% of the time while succeeding on the more complex contradictory inference. This 81.9% gap cannot be explained by task difficulty and instead indicates inconsistent application of logical principles.

Transformation Operations: Inconsistent Handling of Valid Operations. Within transforma-

tion operations, models show inconsistent performance across instances of the same operation type. Claude-3.5-Sonnet achieves 97.7% on I→I conversion but only 26.5% on A→I conversion by limitation. Both are valid conversion operations with similar structural complexity, yet performance differs by 71.2%.

GPT-4.1 shows an even more extreme pattern: 99.7% on I→I conversion versus 2.0% on A→I conversion. This variance cannot be attributed to the inherent difficulty of the logical operation. It suggests that models rely on surface features (e.g., whether the quantifier changes from "All" to "Some") rather than applying consistent conversion rules.

Implications. These patterns indicate that models do not possess stable internal representations of logical operations. Instead, they appear to oscillate between structural reasoning (applying logical rules) and surface pattern matching (responding based on lexical or syntactic cues), with the choice of reasoning mode varying unpredictably across instances. This instability fundamentally limits the reliability of LLM logical reasoning, as users cannot predict which reasoning mode will be activated for any given input.

E.1.4 Finding 3: Models Lack Robust Operator Grounding

Immediate inference fundamentally requires correct interpretation and manipulation of two classes of logical operators: quantifiers (universal "All/No" vs. particular "Some") and quality markers (affirmative vs. negative, including complement terms like "non-P"). Our analysis reveals that models fail to ground these operators in their formal logical definitions, leading to systematic errors when operator manipulation is required.

Quantifier Processing: Universal Premises Outperform Particular Premises. In opposition relations, model performance varies dramatically based on the quantifier type of the premise proposition. Inference from universal premises (A-type "All S are P" or E-type "No S are P") substantially outperforms inference from particular premises (I-type "Some S are P" or O-type "Some S are not P").

Qwen-2.5-72B achieves 99.3% average accuracy on inferences from A-type premises but only 64.0% on inferences from O-type premises (a gap of 35.3%). This asymmetry suggests that models have more robust representations of universal quantification than particular quantification. Llama-3.1-8B exhibits an extreme form of this pattern: it achieves 95.9% average accuracy on inferences from A-type premises but 0.0% on undetermined cases involving I-type and O-type premises. This complete failure indicates that Llama-3.1-8B cannot reason about particular propositions in contexts requiring uncertainty recognition.

The quantifier asymmetry extends to transformation operations. A→I conversion by limitation requires "downgrading" a universal quantifier to a particular one ("All S are P" → "Some P are S"), and this operation shows near-zero accuracy across multiple models: GPT-4.1 (2.0%), Qwen-2.5-72B (0.0%), Llama-3.1-70B (0.0%), Qwen-2.5-7B (0.0%), and Llama-3.1-8B (0.0%). Models appear unable or unwilling to generate outputs with weaker quantificational force than the input, suggesting a bias toward preserving quantifier strength.

Negation Processing: Complement Operations Systematically Fail. On contraposition, open-source models achieve near-zero accuracy on multiple subtypes:

- Qwen-2.5-72B: A→A: 30.3%, E→Inv: 2.0%, I→Inv: 0.0%, O→O: 10.8%
- Llama-3.1-70B: A→A: 22.2%, E→Inv: 0.0%, I→Inv: 58.0%, O→O: 12.0%
- Qwen-2.5-7B: A→A: 3.5%, E→Inv: 0.0%, I→Inv: 0.0%, O→O: 12.0%
- Llama-3.1-8B: A→A: 21.7%, E→Inv: 0.0%, I→Inv: 0.0%, O→O: 11.0%

Even closed-source models struggle with negation: Claude-3.5-Sonnet achieves 0.0% on both contraposition E→Inv and obversion A→E, and Gemini-2.5-Pro—despite 98.4% overall accuracy—drops to 10.0% on contraposition E→Inv.

These patterns indicate that models have not learned robust representations of logical negation and complementation. The concept "non-P" appears to lack stable grounding, causing systematic failures whenever complement manipulation is required.

Bidirectional Judgment Failures. Models exhibit errors in both directions: committing to definite truth values when the logical relationship is undetermined, and abstaining from judgment when valid inferences are available.

Model	Task	Anonymous	Factual	Counterfactual
Large	Contraposition	-5.15%	-2.86%	+2.14%
	Conversion	+13.24%	+15.00%	+10.00%
	Obversion	+10.14%	+10.53%	+7.24%
Medium	Contraposition	+69.12%	+64.29%	+87.14%
	Conversion	+35.29%	+34.29%	+31.43%
	Obversion	+20.27%	+9.21%	+14.47%
Small	Contraposition	+26.47%	+27.14%	+40.00%
	Conversion	+25.00%	+32.86%	+25.71%
	Obversion	+33.78%	+31.58%	+25.00%

Table 11: Prompt-induced accuracy gains stratified by semantic condition. **Anonymization reveals models' reliance on semantic shortcuts.**

In undetermined cases (where the correct response is "cannot be determined"), Llama-3.1-8B achieves 0.0% accuracy on I→A, I→O, O→A, and O→I, indicating that it always commits to a definite truth value rather than recognizing logical indeterminacy. Conversely, GPT-4.1 achieves only 16.0% on the determinate A→I subalternation, suggesting excessive abstention when definitive inference is warranted.

This bidirectional failure pattern indicates that models lack calibrated confidence in their logical inferences, they cannot reliably distinguish cases where logical rules mandate specific conclusions from cases where no valid inference exists.

E.2 Experiment 2: Robustness Analysis

Experimental Setup. We test the effects of semantics and prompts on immediate inference across models. The evaluation metric is accuracy gain relative to the zero-shot baseline. We also quantify the effectiveness of prompt engineering by measuring the *prompt-induced Accuracy Gain*, which represents the maximum performance improvement achieved by advanced prompting strategies (P2 or P3) relative to the baseline zero-shot performance (P1). Formally, for a given model and task, this metric is defined as:

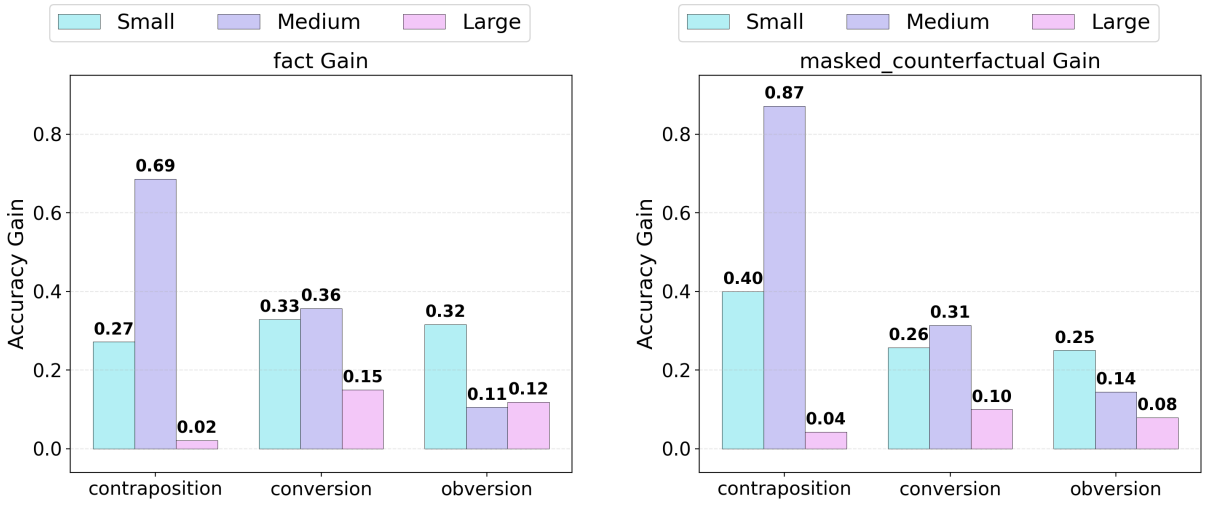
$$\text{Gain} = \max(\text{Acc}_{\text{P2}}, \text{Acc}_{\text{P3}}) - \text{Acc}_{\text{P1}} \quad (1)$$

where $\text{Acc}_{\text{P}i}$ denotes the model's accuracy under prompt format i. This metric captures the potential of logical scaffolding to unlock reasoning capabilities that are latent but inaccessible under standard zero-shot conditions. Table 4 presents results aggregated across semantic conditions, and Table 11 shows gains stratified by semantic condition. Figure 6 shows gains stratified by semantic condition.

Model scale modulates prompt effectiveness. Small models (<10B) benefit from prompting interventions but exhibit instability: contraposition improves substantially with examples (Δ_{Example} = +0.317) while definitions alone show negligible effect (Δ_{Def} = -0.005), yet conversion degrades when examples are added (Δ_{Example} = -0.120) despite strong definition gains (Δ_{Def} = +0.279). Medium models (∼70B) demonstrate the most substantial and consistent gains, achieving +0.736 total improvement on contraposition with both definitions (+0.481) and examples (+0.255) contributing positively; conversion similarly gains +0.298 primarily from definitions (+0.312). Large models show limited improvement across tasks and notably exhibit negative interference on contraposition (Δ_{All} = -0.050), where both definitions (-0.048) and examples (-0.002) degrade performance, suggesting that additional scaffolding disrupts their established reasoning patterns.

Task complexity determines optimal prompting strategy. For operations involving negation (contraposition, obversion), examples (P3) yield greater improvements than definitions alone (P2) for smaller models: small models achieve Δ_{Example} = +0.317 on contraposition and +0.208 on obversion, substantially exceeding their definition gains (-0.005 and +0.093, respectively). For simpler operations (conversion), definitions suffice and additional examples consistently reduce effectiveness across all scales: small models drop by -0.120, medium models by -0.014, and large models by -0.005 when examples are added. This pattern suggests that complex structural transformations involving negation and complement terms require concrete demonstrations to establish correct operator application, whereas simple term-exchange transformations can be triggered by definitional knowledge alone and are disrupted by potentially conflicting examples.

Semantic conditions reveal differential reliance on shortcuts. Removing semantic content amplifies prompt effects and exposes models' varying dependence on surface-level cues. Under counterfactual conditions, medium-scale models achieve the highest contraposition gain (+87.14%), substantially exceeding anonymous (+69.12%) and factual (+64.29%) conditions, suggesting that content conflicting with world knowledge forces greater reliance on structural reasoning. Small models similarly obtain peak contraposition gains under counterfactual conditions (+40.00%



(a) Prompt-induced accuracy gains stratified by factual condition

(b) Prompt-induced accuracy gains stratified by counterfactual condition

Figure 6: Prompt-induced accuracy gains.

vs. +26.47% anonymous and +27.14% factual). However, small models show the opposite pattern for conversion: factual conditions yield the highest gains (+32.86%) compared to anonymous (+25.00%) and counterfactual (+25.71%), indicating continued reliance on semantic scaffolding for simpler operations. Large models exhibit consistent negative effects on contraposition across all semantic conditions (anonymous -5.15%, factual -2.86%, counterfactual +2.14%), with only counterfactual showing marginal positive gains. This stability suggests their negative interference pattern is independent of semantic content and reflects fundamental limitations in how prompting interacts with their internal representations.

E.3 Experiment 3: Failure Mechanism Analysis

E.3.1 Representative Error Examples (One per Error Type)

(Understanding) Over-entailment / Hallucination ($U \to T$).

> **Question:** Given the premise, determine the truth status of the *candidate* statement (`T/F/U`).
> **Premise:** Some non-embargo are not whopper.
> **Candidate:** No non-embargo are whopper.
> **Gold:** Undetermined (`U`).
> **Response:** True (`T`).
> **Reason why wrong:** The premise is existential and only asserts that at least one `non-embargo` is outside the `whopper` set. It does *not* entail the universal exclusion in the candidate (*No non-embargo are whopper*), which would require showing that *all* `non-embargo` instances are *not* `whopper`. The model therefore upgrades an underdetermined case to entailment (over-entailment / hallucination).

(Understanding) Over-contradiction / Forced False ($U \to F$).

> **Question:** Given the premise, determine the truth status of the *candidate* statement (`T/F/U`).
> **Premise:** Some oblong are not square.
> **Candidate:** No oblong are square.
> **Gold:** Undetermined (`U`).
> **Response:** False (`F`).
> **Reason why wrong:** The premise asserts that there exists at least one `oblong` that is *not* `square`, but it does not rule out the possibility that other `oblong` objects *are* `square`. Therefore the universal exclusion in the candidate (*No oblong are square*) is not entailed nor contradicted by the premise, making the correct label `U`. The model incorrectly collapses an underdetermined case into contradiction (forced false).

(Understanding) Over-abstention ($T/F \to U$).

> **Question:** Given the premise, determine the truth status of the *candidate* statement (`T/F/U`).
> **Premise:** All wrong bishop are bishop.
> **Candidate:** Some wrong bishop are bishop.
> **Gold:** True (`T`).
> **Response:** Undetermined (`U`).
> **Reason why wrong:** Under the benchmark's existential-import convention for subalternation, an `A`-form statement (*All S are P*) entails the corresponding `I`-form (*Some S are P*). The candidate is therefore determinate and should be labeled `T`. The model instead abstains, treating a logically licensed inference as undetermined.

(Understanding) Polarity flip ($T \leftrightarrow F$).

> **Question:** Given the premise, determine the truth status of the *candidate* statement (`T/F/U`).
> **Premise:** No non-christmas_card are non-file_transfer_protocol.

Candidate: Some non-christmas_card are not non-file_transfer_protocol.
Gold: True (T).
Response: False (F).
Reason why wrong: The premise is an E-form statement (*No S are P*). Under the benchmark's square-of-opposition convention (existential import), it entails the corresponding O-form subalternation (*Some S are not P*). Here, with S = non-christmas_card and P = non-file_transfer_protocol, the candidate is exactly the entailed O-form and should be labeled T. The model instead predicts F, reversing the polarity of a determinate inference.

(Execution) Quantifier error (QUANTIFIERMISMATCH).

Question: Apply OBVERSION to the premise and output the canonical transformed statement.
Premise: All connected pawns are pawn.
Gold: no connected pawns are non-pawn.
Response: all connected pawns are non-non-pawn.
Reason why wrong: Obversion of an A-form requires switching the quantifier from *All* to the canonical E-form and negating the predicate, yielding *No S are non-P*. The model keeps the *All* quantifier and instead introduces a double negation on the predicate (non-non-pawn), which does not match the required canonical obversion template and is scored as a quantifier mismatch.

(Execution) Negation error (ANYNEGMISMATCH).

Question: Apply OBVERSION to the premise and output the canonical transformed statement.
Premise: Some non-infomercial are not registered_mail.
Gold: some non-infomercial are non-registered_mail.
Response: some infomercial are registered_mail.
Reason why wrong: For an O-form statement (*Some S are not P*), obversion preserves the subject term and converts predicate negation into a complemented predicate, yielding *Some S are non-P*. The response flips *both* negations: it drops the subject prefix (non-infomercial → infomercial) and removes predicate negation (not registered_mail → registered_mail), producing the opposite meaning rather than the required obversion.

(Execution) Term mismatch (ANYTERMMISMATCH).

Question: Apply OBVERSION to the premise and output the canonical transformed statement.
Premise: All connected pawns are pawn.
Gold: no connected pawns are non-pawn.
Response: no non-pawn are connected pawns.
Reason why wrong: Obversion of an A-form (*All S are P*) requires producing the canonical E-form *No S are non-P*, which preserves the subject term and negates the predicate via the complement marker non-. The response instead swaps the argument order (treating non-pawn as the subject) and changes both subject and predicate strings (including pluralization), yielding a different statement form and violating the benchmark's required canonical surface realization. This is therefore annotated as a term mismatch (with co-occurring negation mismatches).

(Execution) Refusal/format error (PARSEERROR_OUTPUT).

Question: Apply CONVERSION to the premise and output the canonical transformed statement.
Premise: All non-ancient_greek are voice.
Gold: some voice are non-ancient_greek.
Response: no valid conversion.
Reason why wrong: The premise is a valid A-form conversion instance under the benchmark rules (conversion per accidens), so the model is required to output the existential converse *Some P are S*. Instead, it produces a refusal/abstention string ("no valid conversion"), which is treated as an invalid-format output and annotated as PARSEERROR_OUTPUT.

(Execution) Over-generation on invalid instances (UNPARSEABLEGOLD).

Question: Apply CONVERSION to the premise; if invalid, output "no valid conversion."
Premise: Some non-andante are not game_law.
Gold: no valid conversion.
Response: some game_law are not non-andante.
Reason why wrong: The premise is an O-form statement (*Some S are not P*). In traditional syllogistic logic, O-form propositions do *not* admit valid conversion, so the correct output is to abstain ("no valid conversion"). The model instead produces a converted sentence, indicating over-generation on an invalid operation.

E.4 Experiment 4: Causal Mechanism Analysis

E.4.1 Causal Pathway Analysis

Experimental Setup. To investigate how model scale affects syllogistic reasoning through intermediate logical capabilities, we employ path analysis with immediate inference as mediating variables. This analysis examines whether scale's effect on syllogistic accuracy operates through foundational logical capabilities (execution and understanding) or through direct pathways that bypass these capabilities.

Variables. We define the following variables:

- *Predictor*: Model scale, encoded as an ordinal variable (0=small, 1=medium, 2=large, 3=large-with-thinking).
- *Mediators*: Execution capability (M_{exec}), computed as the average accuracy across three transformation operations $M_{exec} = (M_{obv} +$

$M_{conv} + M_{contra})/3$. Understanding capability (M_{under}), measured by opposition relation accuracy.

- *Outcome*: Syllogistic reasoning accuracy (Y), evaluated on categorical syllogisms.

- *Covariate*: Prompt condition (0=zero-shot, 1=few-shot), included to control for prompting effects.

Structural Equations. The path model specifies the following relationships:

$$\begin{aligned} M_{exec} &= \alpha_1 + a_1 \cdot \text{Scale} + a_3 \cdot \text{Prompt} + e_1, \\ M_{under} &= \alpha_2 + a_2 \cdot \text{Scale} + a_4 \cdot \text{Prompt} + e_2, \\ Y &= \alpha_3 + b_1 \cdot M_{exec} + b_2 \cdot M_{under} \\ &\quad + c' \cdot \text{Scale} + c'' \cdot \text{Prompt} + e_3, \end{aligned} \tag{2}$$

where a_1 and a_2 quantify scale's effect on intermediate capabilities, b_1 and b_2 measure the contribution of each mediator to syllogistic accuracy, and c' captures the direct effect of scale that bypasses the measured logical capabilities. The indirect effects through execution and understanding are computed as $a_1 \times b_1$ and $a_2 \times b_2$, respectively.

Semantic Conditions. To disentangle logical reasoning from semantic priors, analyses are conducted separately for three semantic conditions (factual, anonymous, counterfactual) with $N = 16$ observations per condition.

Sample and Limitations. Each condition includes $N = 16$ observations (8 models $\times$ 2 prompt conditions), yielding limited statistical power. Given this constraint, we interpret results as exploratory, emphasizing effect size patterns over p-values.

Our path analysis results are shown in Table 13.

Residual Variance Analysis. The residual standard deviations ($\hat{\sigma}_{e1}, \hat{\sigma}_{e2}, \hat{\sigma}_{e3}$) quantify unexplained variance after accounting for predictors and mediators (Table 12). Crucially, $\hat{\sigma}_{e3}$ is substantially lower in the anonymous condition (0.018) than in factual (0.048) or counterfactual (0.040) conditions. This indicates that when semantic cues are removed, logical mediators largely capture scale's effect on reasoning. In contrast, the larger $\hat{\sigma}_{e3}$ in the factual condition suggests unmeasured semantic shortcuts contribute substantially to performance.

Table 13 presents path coefficients across semantic conditions. We discuss each condition in detail below.

Table 12: Residual Standard Deviations by Semantic Condition

Residual Term	Factual	Anonymous	Counterfactual
$\hat{\sigma}_{e1}$ (M_{exec})	0.092	0.131	0.115
$\hat{\sigma}_{e2}$ (M_{under})	0.084	0.065	0.058
$\hat{\sigma}_{e3}$ (Y)	0.048	0.018	0.040

Finding 1: Semantic condition determines whether understanding contributes to reasoning. The contribution of understanding capability (b_2) to syllogistic accuracy exhibits strong condition dependence. In the factual condition, $b_2 = 0.145$ ($p = 0.517$) is non-significant; in the counterfactual condition, $b_2 = 0.949$ ($p = 0.018$) reaches significance; in the anonymous condition, $b_2 = 1.306$ ($p < 0.001$) is highly significant. The effect size increases approximately 9-fold from factual to anonymous, indicating that the removal of semantic cues "activates" the predictive role of understanding capability. When models cannot rely on world knowledge, opposition relation comprehension becomes the decisive factor for syllogistic performance.

Finding 2: Execution capability shows no independent predictive contribution across all conditions. In contrast to understanding, execution capability (b_1) fails to reach statistical significance in any semantic condition: factual $b_1 = 0.324$ ($p = 0.141$), anonymous $b_1 = -0.058$ ($p = 0.507$), counterfactual $b_1 = 0.054$ ($p = 0.736$). Once understanding capability is controlled for, transformation operation execution no longer contributes additional variance to syllogistic accuracy. Opposition relation comprehension serves as the critical mediating capability linking immediate inference to syllogistic reasoning.

Finding 3: Logical pathways are bypassed by semantic shortcuts under factual conditions. In the Factual condition, although scale significantly enhances execution ($a_1 = 0.204$, $p < 0.001$) and understanding ($a_2 = 0.049$, $p = 0.050$) capabilities, neither capability significantly contributes to syllogistic accuracy (b_1 and b_2 both non-significant), and the direct effect of scale is also non-significant ($c' = 0.042$, $p = 0.306$). Combined with the highest residual variance in this condition ($\hat{\sigma}_{e3} = 0.048$), this pattern suggests that models bypass logical pathways when semantic cues are available. Consider the syllogism: "All doctors are humans; All humans are mortal; Therefore, all doctors are mortal." Models can derive the answer directly from world knowledge without

Table 13: Path Analysis Results with Statistical Significance.

	Factual		Anonymous		Counterfactual	
Path	Coef	p	Coef	p	Coef	p
a_1: Scale $\rightarrow$ Execution	0.204	<0.001	0.226	<0.001	0.196	<0.001
a_2: Scale $\rightarrow$ Understanding	0.049	0.050	0.050	0.016	0.043	0.018
b_1: Execution $\rightarrow Y$	0.324	0.141	-0.058	0.507	0.054	0.736
b_2: Understanding $\rightarrow Y$	0.145	0.517	1.306	<0.001	0.949	0.018
c': Scale $\rightarrow Y$ (direct)	0.042	0.306	0.097	<0.001	0.074	0.025

executing the logical reasoning chain.

Finding 4: Counterfactual condition exhibits hybrid characteristics. The path coefficient pattern in the counterfactual condition falls between factual and anonymous: understanding is significant but with a weaker effect ($b_2 = 0.949$ vs. 1.306 in anonymous), the direct effect is significant but smaller ($c' = 0.074$ vs. 0.097 in anonymous), and residual variance is intermediate ($\hat{\sigma}_{e3} = 0.040$). This reflects the nature of counterfactual reasoning—models must employ logical capabilities to process formal structure (e.g., "All unicorns are purple; All purple things are magical; Therefore, all unicorns are magical"), but the fictional entities in the premises still provide a semantic plausibility structure that allows some non-logical factors to participate in the reasoning process.

E.4.2 Conditional Probability

Experimental Setup. To analyze the causal influence of immediate inference on syllogistic reasoning, we design a conditional probability experiment. Specifically, we apply logically equivalent transformations (conversion, obversion, contraposition) to syllogism premises and evaluate whether models can still derive valid conclusions. We compute three conditional probability metrics to quantify the dependency between immediate inference (M) and syllogistic reasoning (Y):

- **TRC** (Causal Sufficiency): $P(Y = 1 \mid M = 1)$. If the foundational logic is correct, is the syllogism necessarily correct? Measures the *effectiveness of the logical pathway*. Higher values indicate more robust logical chains.

- **FI** (Causal Necessity): $P(Y = 0 \mid M = 0)$. If the foundational logic is wrong, is the syllogism necessarily wrong? Measures the *irreplaceability of the logical pathway*. Lower values indicate the existence of shortcuts.

- **SCR** (Semantic Compensation Rate): $P(M = 0 \mid Y = 1)$. Answered the syllogism correctly but failed the foundational transformation? Measures *illusory competence*. Higher values indicate greater reliance on semantic shortcuts rather than logical derivation.

Results are shown in Figure 7, Figure 8 and Figure 9.

Immediate Inference as Foundation. TRC exhibits consistent scale effects across all operations: large models achieve TRC = 0.74-0.98, medium models 0.54-0.97, and small models 0.29-0.72. This pattern indicates that complex transformations more effectively differentiate models' genuine logical reasoning capabilities. Simple operations yield similar performance across scales, while complex operations substantially amplify scale differences.

F Cross-Task Generalization and Model Training

F.1 Cross-Task Generalization

Experimental Setup. To examine whether immediate inference capabilities generalize to broader reasoning tasks, we compute correlation coefficients between immediate inference performance and seven downstream reasoning benchmarks.

Datasets Selection. We select benchmarks spanning diverse reasoning paradigms. **ControlNLI** (Liu et al., 2021a) evaluates natural language inference over long contexts, requiring logical and semantic consistency judgments. **ProofWriter** (Tafjord et al., 2021) tests multi-step deductive reasoning through rule derivation and chain reasoning. **RuleTaker** (Clark et al., 2020) examines natural language rule-based inference at varying reasoning depths. **LogicBench-PL** (Parmar et al., 2024) assesses propositional logic derivation with emphasis on negation han-

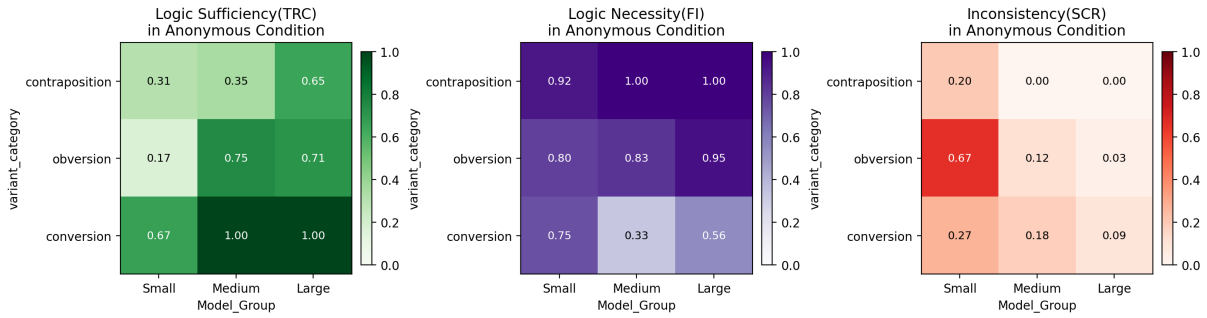


Figure 7: Conditional probability metrics under anonymous condition.

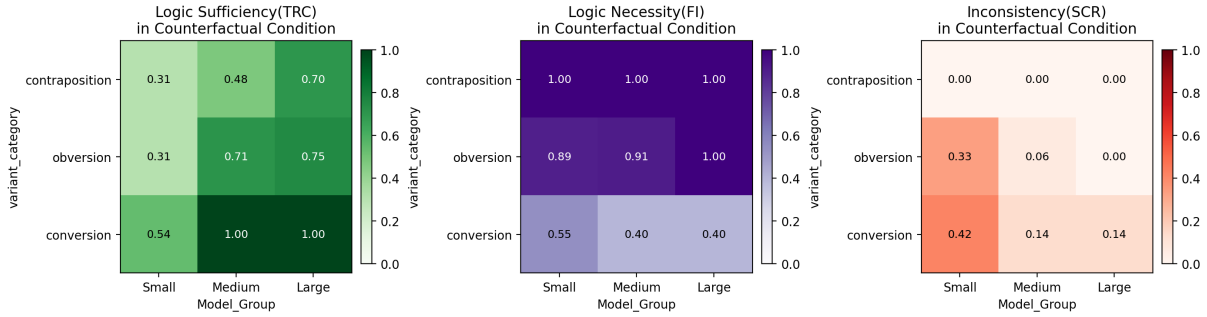


Figure 8: Conditional probability metrics under counterfactual condition.

dling. **LogicBench-FOL** (Parmar et al., 2024) targets first-order logic with quantifier reasoning and structured inference. **LogicBench-NM** evaluates non-monotonic reasoning involving defeasible conclusions that can be retracted given new information (Parmar et al., 2024). Finally, **LogiQA** (Liu et al., 2021b) evaluates exam-style multiple-choice logical reasoning in machine reading comprehension, emphasizing conditional and case-based inference over alternative scenarios, which is closely related to modal/conditional-logic style reasoning. We randomly select 300 test instances from each benchmark, and the results are shown in Table 14.

We compute three correlation coefficients: Pearson (Pearson, 1895), Spearman (Spearman, 1904), and Kendall (Kendall, 1938) between each immediate inference capability metric and each downstream benchmark. Results are shown in Figure 10.

Beyond the positive/negative correlation patterns reported in Section 5.5, two additional observations emerge from the full correlation matrix:

Finding 1: Moderate Correlations with Rule-Based and First-Order Logic Benchmarks. Correlations with RuleTaker ($\rho = 0.67$–0.80) and LogicBench-FOL ($\rho = 0.55$–0.79) are moderate but positive. The relatively lower correlation with LogicBench-FOL may reflect that first-order logic involves additional complexities (polyadic predicates, nested quantifiers) beyond the monadic predicates in immediate inference. RuleTaker's moderate correlation suggests that natural language rule-following draws on immediate inference capabilities but also involves additional linguistic processing.

Finding 2: Execution Metrics Outperform Understanding Metrics. Transformation operation metrics (Conversion, Obversion, Contraposition) consistently show stronger correlations with downstream benchmarks than opposition relation metrics (Understanding). For ConTRoL, Spearman correlations are: conversion $\rho = 0.95$, obversion $\rho = 0.86$, contraposition $\rho = 0.95$, versus understanding $\rho = 0.67$. Similar patterns hold across other benchmarks.

This asymmetry suggests that the ability to *execute* logical transformations is more predictive of downstream reasoning performance than the ability to *understand* truth-value relationships. Execution requires generating correct outputs under structural constraints, which may better reflect the generative reasoning demands of downstream tasks.

These additional observations complement the main-paper findings (Section 5.5) in two ways: (1) correlations with rule-based (RuleTaker) and first-order logic (LogicBench-FOL) benchmarks are positive but only moderate, indicating that imme-

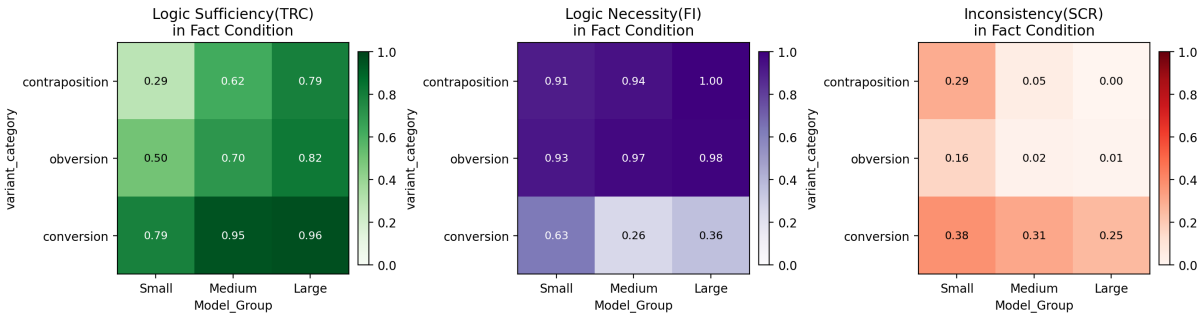


Figure 9: Conditional probability metrics under factual condition.

Model	ConTRoL *NLI*	ProofWriter *Deduction*	RuleTaker *Rule-based*	LogicBench *(PL)*	LogicBench *(FOL)*	LogicBench *(NM)*	LogiQA *MCQ*
Closed-Source Models							
Gemini-2.5-Pro	**82.0**	96.7	**74.0**	**89.3**	**79.0**	58.3	**84.7**
GPT-o3	79.7	**99.7**	**74.0**	84.7	78.0	61.0	82.7
GPT-4.1	69.7	83.3	68.0	76.3	78.0	63.0	63.3
Claude-3.5-Sonnet	78.7	85.3	67.3	80.0	75.7	59.7	63.7
Open-Source Models							
Qwen-2.5-72B	61.7	72.3	65.0	76.7	77.0	**67.3**	60.3
Llama-3.1-70B	65.0	70.3	72.0	76.0	78.3	63.0	55.7
Qwen-2.5-7B	55.7	53.3	66.0	69.7	70.0	62.0	55.3
Llama-3.1-8B	50.3	47.0	65.0	67.7	70.7	65.3	44.3

Table 14: Model performance (%) on downstream reasoning benchmarks. Best results are **bolded**, second-best is underlined.

diate inference captures a core but not exhaustive component of these more linguistically or quantificationally complex tasks; and (2) transformation operation metrics (execution) are consistently more diagnostic of downstream reasoning than opposition relation metrics (understanding), suggesting that the ability to generate correct outputs under structural constraints better predicts downstream performance than declarative knowledge of truth-value relationships.

F.2 Model Training

Goal and controlled design. We conduct a controlled training intervention study to test whether *Immediate Inference* (II) serves as a foundational capability that transfers to downstream reasoning benchmarks. We use **Qwen-2.5-7B** and **Llama-3.1-8B** as the base models and construct *training-only* corpora for II and syllogism using the same seed construction and validation pipeline as Appendix D (i.e., the same rule-based generation procedure and three-stage quality control), but with a **disjoint seed set** from the benchmark evaluation pairs. To ensure strict separation from the evaluation benchmark and prevent near-duplicate memorization, we enforce that **each training instance differs from the test pairs by at least one term** (subject, predicate, or middle term), thereby guaranteeing measurable discrepancies between training and test distributions.

Most importantly, the total supervision volume is **strictly matched** across the **II-Only**, **Syllogism-Only**, and **FOL-Only** conditions.

For **Syllogism-Only**, we construct training data from the **24 valid syllogistic forms**. For each form, we generate **4 groups** following a semantic ratio of **Fact : Anonymous : Counterfactual = 2 : 1 : 1**, and each group contains **8 concrete instances**. This yields a total of **768 training instances**:

$$24 \text{ forms} \times 4 \text{ groups} \times 8 \text{ instances} = 768.$$

For **II-Only**, we strictly down-sample and balance the II corpus to match this supervision volume exactly. Across the three core transformation modes—**Obversion**, **Conversion**, and **Contraposition**—we uniformly sample **256 instances per mode**, yielding **768 training instances**:

$$3 \text{ modes} \times 256 \text{ instances} = 768.$$

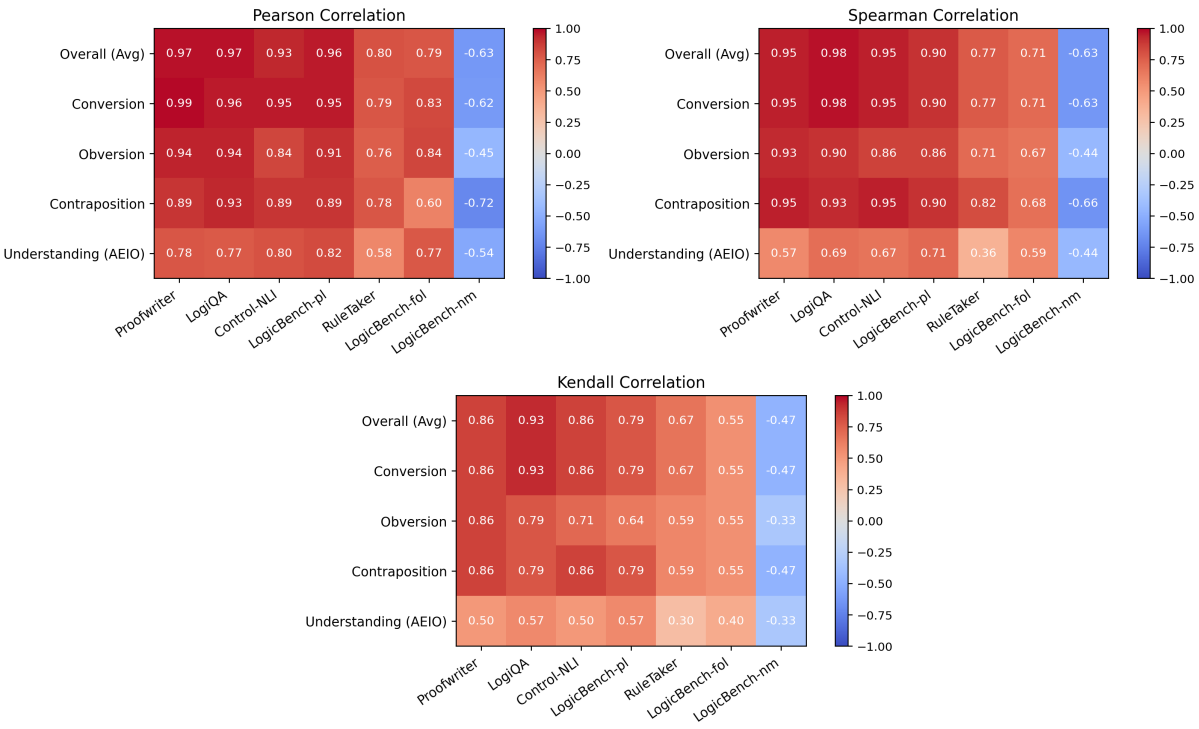


Figure 10: Correlation heatmaps between immediate inference capabilities and downstream reasoning benchmarks.

For **FOL-Only**, to align with the **II-Only** and **Syllogism-Only** settings, we randomly sampled **768 training instances** from the LogicBench first-order logic training set, yielding the same total supervision volume.

Thus, the **II-Only**, **Syllogism-Only**, and **FOL-Only** conditions are matched exactly in training size (**768** instances each), ensuring a fair comparison of transfer effects.

Training conditions. We compare the following intervention conditions:

1. **Baseline**: no further training (zero-shot evaluation of the base model).
2. **Syllogism-Only**: fine-tuning exclusively on the **768** syllogistic training instances.
3. **II-Only**: fine-tuning exclusively on the **768** II training instances.
4. **FOL-Only**: fine-tuning exclusively on the **768**-instance LogicBench first-order logic training set.
5. **Mixed-Full**: fine-tuning on both full II and syllogistic training corpora.
6. **Mixed-Reduced**: fine-tuning on the full II corpus together with **50%** of the syllogistic data used in **Mixed-Full**, to test whether additional syllogistic supervision provides incremental benefit beyond II.

F.3 Chain-of-Thought Setting

Experimental Setup. To test whether our findings depend on answer-only prompting, we evaluate a CoT-style variant of the benchmark by prepending an internal chain-of-thought instruction to each task-specific prompt (Wei et al., 2022), while keeping the output format, normalization rules, and correctness criteria unchanged. Models are encouraged to reason step by step internally, but are still evaluated only on their final structured prediction.

We compare the CoT and answer-only settings on the same examples for the same model and task, measuring two types of similarity. **Effect similarity** compares correctness outcomes using agreement and Cohen's κ (Cohen, 1960). **Output similarity** compares normalized final responses using exact match, normalized Levenshtein similarity (Levenshtein et al., 1966), token-level F1, and ROUGE-L F1 (Lin, 2004). Results are shown in Table 15.

Prompt Format: Internal chain-of-thought prefix

```
``Think step by step internally to ensure
    correctness, but DO NOT write your
    reasoning. Only output exactly what is
    requested.

[Then append the same task-specific prompt as
    in Format 1.]''
```

Aggregate Stability Under CoT. At the aggregate level, semantic behavior is highly stable with and without CoT. Agreement remains high across all four task families, ranging from 0.860 to 0.925, with substantial consistency as measured by Cohen's κ (0.771–0.838). Accuracy differences are also small: AEIO changes only from 0.789 to 0.796, conversion is essentially unchanged (0.595 vs. 0.594), obversion differs only marginally (0.596 vs. 0.598), and contraposition shows the largest but still modest change (0.433 vs. 0.425). Adding an internal CoT prefix therefore does not substantially alter final semantic predictions, indicating that immediate-inference deficits are not easily resolved by eliciting additional reasoning.

Structural Complexity Modulates CoT Sensitivity. Semantic consistency is strongest on the more template-constrained tasks. Conversion exhibits the highest stability (agreement 0.925, exact match 0.925, Levenshtein 0.973, token F1 0.925); obversion is similarly stable (agreement 0.882, exact match 0.882, Levenshtein 0.943, token F1 0.889); AEIO also remains robust (agreement 0.893, $\kappa = 0.817$). By contrast, contraposition is the least stable setting, with the lowest agreement (0.860), lowest κ (0.771), and lowest similarity scores overall (Levenshtein 0.919; token F1 0.865). This pattern is consistent with contraposition's structural demands, which require coordinated manipulation of polarity, complementation, and term order.

Consistency Does Not Imply Correctness. The per-model results show that consistency and correctness are distinct properties. On contraposition, the weakest models (Qwen-2.5-7B, Llama-3.1-8B, Qwen-2.5-72B, Llama-3.1-70B) all exhibit agreement above 0.98 while accuracy remains at 0.07–0.19 in both settings—systematically wrong, but stably so. Conversely, ceiling-saturated models (Gemini-2.5-Pro, GPT-o3) are stable because they are nearly always correct (accuracy ≥ 0.95, agreement ≥ 0.93). The largest CoT-induced shifts cluster on models operating away from both regimes: GPT-4.1 on contraposition exhibits the lowest agreement in its task block (0.886) and a 6.6-point accuracy drop (0.427 $\rightarrow$ 0.361), with a similar though milder pattern on obversion (agreement 0.864). CoT therefore affects predictions most when models are genuinely uncertain, not when they are strong; stable outputs across conditions are compatible with either high-accuracy competence or low-accuracy rote behavior, and cannot be read as evidence of sound underlying reasoning.

Analysis of the CoT setting yields three findings: (1) aggregate semantic behavior is highly stable with and without CoT, with agreement above 0.85 and near-identical accuracies across all four task families, indicating that immediate-inference deficits are not artifacts of answer-only prompting; (2) CoT sensitivity scales with the structural complexity of the transformation, with contraposition showing both the lowest output agreement and the largest individual CoT-induced accuracy shift (GPT-4.1, -6.6 points); and (3) output consistency across conditions is not a reliable proxy for reasoning quality, as weaker models can remain highly consistent while systematically wrong.

Task	# Models	Agreement	Cohen's κ	Exact Match	Lev. Sim.	Token F1	Acc. CoT	Acc. Base
AEIO	8	0.893	0.817	0.893	0.927	0.893	0.796	0.789
Conversion	8	0.925	0.838	0.925	0.973	0.925	0.594	0.595
Obversion	8	0.882	0.798	0.882	0.943	0.889	0.598	0.596
Contraposition	8	0.860	0.771	0.860	0.919	0.865	0.425	0.433

Table 15: Semantic-output similarity statistics between the internal chain-of-thought setting (COT) and the baseline setting (BASE), aggregated across all compared models for each task family. Agreement and Cohen's κ are computed over final predictions; Exact Match is the proportion of identical outputs; Lev. Sim. denotes average normalized Levenshtein similarity; Token F1 is the average token-level overlap. Accuracies are reported separately for the CoT and Base settings.

Task	Model	Agreement	Cohen's κ	Exact Match	Lev. Sim.	Token F1	Acc. CoT	Acc. Base
AEIO	Gemini-2.5-Pro	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.998**
	GPT-o3	0.858	0.594	0.858	0.876	0.858	0.778	0.771
	GPT-4.1	0.938	0.830	0.931	0.939	0.931	0.758	0.764
	Claude-3.5-Sonnet	0.974	0.891	0.968	0.972	0.968	0.857	0.862
	Qwen-2.5-72B	0.991	0.966	0.991	0.992	0.991	0.845	0.842
	Llama-3.1-70B	0.972	0.910	0.969	0.972	0.969	0.798	0.812
	Qwen-2.5-7B	0.994	0.985	0.993	0.993	0.993	0.706	0.711
	Llama-3.1-8B	0.986	0.972	0.978	0.983	0.978	0.565	0.556
Conversion	Gemini-2.5-Pro	**0.998**	0.000	0.913	0.981	0.977	**1.000**	**0.998**
	GPT-o3	0.982	0.068	0.978	**0.992**	**0.989**	0.992	0.988
	GPT-4.1	0.982	0.959	0.969	0.982	0.976	0.679	0.688
	Claude-3.5-Sonnet	0.965	0.920	0.954	0.977	0.969	0.680	0.682
	Qwen-2.5-72B	0.993	0.986	**0.985**	0.990	**0.989**	0.512	0.517
	Llama-3.1-70B	0.994	**0.988**	0.965	0.982	0.978	0.499	0.501
	Qwen-2.5-7B	0.986	0.961	0.962	0.981	0.980	0.235	0.232
	Llama-3.1-8B	0.978	0.919	0.928	0.973	0.972	0.171	0.158
Obversion	Gemini-2.5-Pro	**0.995**	-0.002	0.870	0.967	0.966	0.997	**0.998**
	GPT-o3	0.993	0.425	0.865	0.984	0.965	**0.995**	0.993
	GPT-4.1	0.864	0.673	0.793	0.978	0.945	0.703	0.710
	Claude-3.5-Sonnet	0.973	0.947	0.932	0.986	0.977	0.497	0.503
	Qwen-2.5-72B	0.978	0.952	**0.953**	**0.993**	**0.987**	0.623	0.620
	Llama-3.1-70B	0.973	0.947	0.935	0.981	0.978	0.482	0.488
	Qwen-2.5-7B	0.983	**0.958**	0.944	0.983	0.980	0.268	0.268
	Llama-3.1-8B	0.973	0.913	0.858	0.958	0.940	0.207	0.186
Contraposition	Gemini-2.5-Pro	0.979	0.754	0.799	0.963	0.933	0.954	**0.958**
	GPT-o3	0.926	0.002	0.748	0.938	0.903	**0.966**	0.957
	GPT-4.1	0.886	0.763	0.758	0.900	0.854	0.361	0.427
	Claude-3.5-Sonnet	0.955	0.910	0.914	0.987	0.976	0.528	0.532
	Qwen-2.5-72B	0.985	**0.952**	**0.964**	**0.990**	**0.987**	0.188	0.191
	Llama-3.1-70B	0.981	0.936	0.938	**0.990**	0.980	0.185	0.180
	Qwen-2.5-7B	**0.992**	0.942	0.958	0.984	0.979	0.072	0.072
	Llama-3.1-8B	0.986	0.945	0.935	0.980	0.978	0.148	0.151

Table 16: Per-model CoT vs. Base semantic-output similarity and accuracy across four task families. Sample sizes: AEIO (n=1100), Conversion, Contraposition (n=1300), Obversion (n=1200). Per task and column, best values are in **bold** and second-best are underlined.