# Understanding Syllogistic Reasoning in LLMs from Formal and Natural Language Perspectives

**Aheli Poddar**[1], **Saptarshi Sahoo**[2], **Sujata Ghosh**[2]

[1]Institute of Engineering & Management, Kolkata
[2]Indian Statistical Institute, Chennai
aheli.poddar2022@iem.edu.in, saptarshi@isichennai.res.in, sujata@isichennai.res.in

**Abstract**

We study syllogistic reasoning in LLMs from the logical and natural language perspectives. In process, we explore fundamental reasoning capabilities of the LLMs and the direction this research is moving forward. To aid in our studies, we use 14 large language models and investigate their syllogistic reasoning capabilities in terms of symbolic inferences as well as natural language understanding. Even though this reasoning mechanism is not a uniform emergent property across LLMs, the perfect symbolic performances in certain models make us wonder whether LLMs are becoming more and more formal reasoning mechanisms, rather than making explicit the nuances of human reasoning.

**Code** — https://github.com/XAheli/Logic-in-LLMs

## 1 Introduction

With the unprecedented development of large language models (LLMs) in recent years that have made them resemble human speakers and reasoners to a great extent in many levels (Holliday, Mandelkern, and Zhang 2024; Bubeck et al. 2023; Zhao et al. 2023), the reasoning capabilities of LLMs have increased manifold. To motivate such growth, the question we generally ask an LLM is to what extent the LLM has grasped logical reasoning in its different forms, for example, see (Holliday, Mandelkern, and Zhang 2024; Borazjanizadeh and Piantadosi 2024; Sambrotta 2025). In contrast, the motivation for this study is somewhat distinct in nature in that we wonder whether developing LLM to have excellent logical reasoning capabilities is fruitful in the long run, as having such features does not bring an LLM closer to mimicking human reasoning. As a case in point, we consider syllogistic reasoning from a formal as well as natural language viewpoint.

Evidently, humans are far from logical when it comes to reasoning, and they are often influenced by their past experiences and knowledge, for example, consider the belief-bias effect (Evans, Barston, and Pollard 1983): People doing syllogistic reasoning are often influenced by the believability of the conclusion. In fact, it is shown by (Lewton 2016) that individuals with autistic traits show less belief-bias effect than typical individuals. In this scenario, one might consider to check whether LLM reasoning is close to human reasoning by studying the belief-bias effect on the LLMs, and the present work studies this question. We note that (Eisape et al. 2024) studied a similar question, but their methodology is quite different from ours. Before describing the exact contribution of this work, let us discuss some recent work on syllogistic reasoning in LLMs.

A novel framework dealing with legal syllogistic reasoning is provided in (Zhang et al. 2025). In this work, the LLMs are empowered to provide explicit and trustworthy legal reasoning by integrating a retrieval mechanism with reinforcement learning. A mechanistic interpretation of syllogistic reasoning is provided in (Kim, Valentino, and Freitas 2025). This work deals with belief-biases as well and it is shown that such biases contaminate the reasoning mechanisms. In (Zong and Lin 2024), the authors make a detailed survey on the reasoning capabilities of LLMs with respect to categorical syllogisms.

This work makes several key contributions to understanding syllogistic reasoning in LLMs from both formal and natural language perspectives. We introduce a novel dual ground truth framework that evaluates each syllogism on two separate dimensions: syntactic validity *(does the conclusion logically follow?)* and natural language believability *(is the conclusion intuitively plausible?)*. These two dimensions may align or conflict with each other, enabling us to assess formal reasoning capabilities independently from natural language understanding. Through a comprehensive empirical study, we systematically evaluated 14 state-of-the-art LLMs across four prompting strategies and three temperature settings on carefully constructed syllogisms covering diverse logical structures and belief-bias conditions. Our analysis reveals that the majority of models exhibit a significant measure of belief bias; in other words, they perform better on certain kinds of problem (where logic aligns with intuition) than others. We further uncover a substantial gap between syntactic and natural language understanding accuracy, demonstrating that current LLMs excel at formal logical structure while struggling with natural language plausibility judgments—a pattern opposite to human reasoning tendencies. Contrary to conventional wisdom, we find that few-shot prompting degrades performance compared to zero-shot, and that reasoning capability depends critically on

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| B-C<br>C-D | C-B<br>D-C | B-C<br>D-C | C-B<br>C-D |

Table 1: A description of the four figures for syllogisms containing the variables B, C, and D.

architectural choices rather than raw parameter count. These findings raise a fundamental question: Are LLMs evolving into formal reasoning engines that surpass human-like reasoning with its inherent biases?

The remainder of the paper is structured as follows. Section §2 provides a brief overview of syllogisms. Section §3 delves into the experimental details, including the models, data, overall methodology, prompting variants, and evaluation metrics. Section §4 reports on the findings and their interpretations. Section §6 provides a discussion of the limitations of our study, and Section §7 concludes the article.

## 2 On Syllogisms

The concept of *syllogism* was first introduced by Aristotle (Smith et al. 1989), and as observed by Robin Smith (Smith 2022), a syllogism in modern logic consists of three subject-predicate propositions, two premises, and a conclusion, and whether or not the conclusion follows from the premises. An example of syllogism is as follows: *"No footballer is a swimmer; Some swimmers are gardeners; Therefore, some gardeners are not footballers."* When terms like *footballer* or *swimmer* are replaced by generic terms like B, C and D, we can rewrite the above premises by: *"No B is C; Some C are D."* A conclusion relates the non-shared terms, for example, *"Some D are not B"*.

In the literature, various types of syllogisms are studied, categorical, conditional, and others (Copi, Cohen, and McMahon 2016). In this work, we mostly concentrate on categorical syllogisms, but we consider a few others as well. The statements of a categorical syllogism look like the following: *Quantifier (Subject) Copula (Predicate),* which take four standard forms, viz.

- *Universal Affirmative (A)*: All S are P, i.e., $S \subseteq P$.
- *Universal Negative (E)*: No S is P, i.e., $S \cap P = \emptyset$.
- *Particular Affirmative (I)*: Some S is P, i.e., $S \cap P \neq \emptyset$.
- *Particular Negative (O)*: Some S is not P, i.e., $S \setminus P \neq \emptyset$.

Here, S is the subject and P is the predicate. S and P are generally termed variables, and these quantifier styles, namely, A, E, I, O, are called 'moods'. The variables may change their orders, leading to new premises. As mentioned earlier, one of the three variables used in a syllogism is not there in the conclusion, and evidently the variable is common to both premises. Depending on the placement of the common variable (C, say) that does not occur in the conclusion, we get four types of *figures* for syllogisms. See Table 1 for a detailed description.

We should note here that, in statements of type A, 'All' is sometimes overlooked for the sake of simplicity. The following example clarifies the point: *"All vehicles have wheels; Boats are vehicles / A boat is a vehicle; Therefore, boats have wheels / a boat has wheels."*

A syllogism is said to be *valid* if the truth of the premises implies the truth of the conclusion. A way to check the validity of a syllogism is by converting the statements in a suitable first order language and check the validity there. The other way is through enumerating each case (there will be some finite number of cases where the two premises will have one of the four forms A, E, I or O) and then using standard Venn Diagram techniques to fix the conclusion. Thus, when a new tuple of syllogism comes in, the job of checking validity boils down to just checking the instance from the already defined cases and to conclude from it.

A syllogism is said to be *believable* if the conclusion of the syllogism is actually true. For this case, the logical argument does not play any role. The main goal of this research work is two-fold. On one hand, we would like to check how accurately the LLMs can do syllogistic reasoning, and on the other hand we would like to check whether context and real world knowledge play any role in their reasoning processes. To this end, the following four categories of syllogisms play a significant role, namely (i) valid-believable, (ii) valid-unbelievable, (iii) invalid-believable, and (iv) invalid-unbelievable. These distinct types are summarized in Table 2, given in (Braüner, Ghosh, and Ghosh 2025), which provides an example for each such type of syllogism.

## 3 Experiments

We conduct a systematic evaluation of syllogistic reasoning capabilities across diverse language models, examining the effects of prompting strategies, temperature settings, and content variations on logical inference accuracy. Our experimental design encompasses 168 unique configurations (14 models $\times$ 4 strategies $\times$ 3 temperatures), enabling comprehensive analysis of factors influencing LLM syllogistic reasoning performance.

### 3.1 Models

We evaluated syllogistic reasoning capabilities in 14 large language models spanning 8 organizations, listed in Table 3. The Google Gemini models were accessed through Google AI Studio APIs.[1] All remaining models were accessed via the HuggingFace Inference API[2] using the *:cheapest* routing for automatic provider selection.[3] Our model selection prioritized four criteria: (1) organizational diversity to capture different development philosophies, (2) parameter scale range (1B to 671B) to assess scaling effects, (3) architectural variety including dense transformers and Mixture-of-Experts (MoE) systems, and (4) API reproducibility.

### 3.2 Data and Methodology

**Dataset Construction** For our experiments, we constructed a benchmark of 160 syllogisms, mostly categorical, adapted from the cognitive science and psychology literature on human syllogistic reasoning (Solcz 2008; Lewton 2016).

[1] https://ai.google.dev/gemini-api/docs

[2] https://huggingface.co/docs

[3] Total API costs for all experiments were approximately $ 500

| | Believable | Unbelievable |
|---|---|---|
| **Valid** | *All birds have feathers*<br>*Robins are birds*<br>*Therefore robins have feathers* | *All mammals walk*<br>*Whales are mammals*<br>*Therefore whales walk* |
| **Invalid** | *All flowers need water*<br>*Roses need water*<br>*Therefore roses are flowers* | *All insects need oxygen*<br>*Mice need oxygen*<br>*Therefore mice are insects* |

Table 2: Example syllogisms illustrating the four categories described in §2.

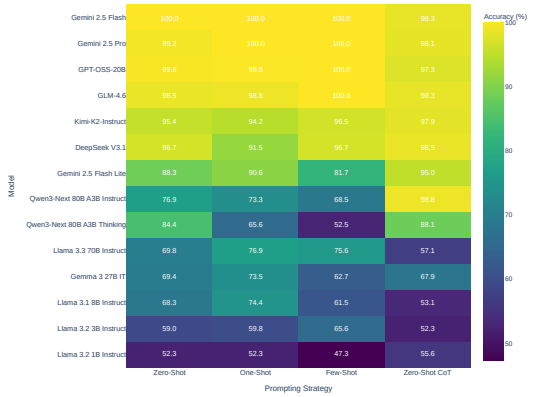


Figure 1: Heatmap of model accuracy across four prompting strategies (Zero-shot, One-shot, Few-shot, Zero-shot Chain-of-Thought). Despite few-shot showing significant mean decline ($\Delta = -3.57$ pp, $p = 0.0165^*$), systematic patterns across models remain minimal, indicating strategy effects are model-specific rather than universal.

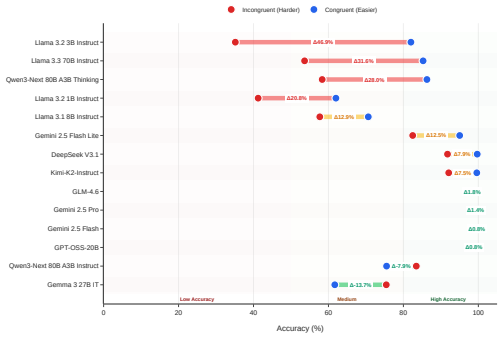


Figure 2: Belief bias effect across 14 models comparing performance on congruent syllogisms (logic aligns with intuition) versus incongruent syllogisms (logic conflicts with intuition). Twelve models (86%) exhibit positive bias ($\Delta = +10.81$ pp, $p = 0.0280^*$, $d = 0.66$). Top-tier models show minimal bias ($< 2$ pp), while lower-tier models show severe bias (up to $+46.9$ pp). Negative correlation ($\rho = -0.565^*$) indicates higher reasoning ability reduces reliance on semantic heuristics.

We began with 40 base syllogisms, each handcrafted to cover different syllogistic figures and validity conditions. To isolate the effects of logical structure from natural language-content, given our *dual ground truth annotations*, we created three additional variants for each base syllogism. The *nonsense variant* (X) replaces meaningful predicates with abstract terms (e.g., *"blargs"*, *"zimons"*, *"glorps"*), testing pure logical reasoning without natural language interference. The *order-switched variant* (O) reverses the order of presentation of the premises to test the sensitivity to the structure of the argument. The *combined variant* (OX) applies both modifications, providing a comprehensive robustness assessment.

For example, the normal variant *"All calculators are machines; All computers are calculators; Therefore, some machines are not computers"* becomes *"All blargs are zimons; All glorps are blargs; Therefore, some zimons are not glorps"* in its nonsense form. We reviewed all stimuli and made necessary adjustments by hand to ensure grammatical correctness and logical equivalence across variants.

**Dual Ground Truth** Each syllogism carries two independent ground truth annotations, enabling orthogonal evaluation of logical reasoning and natural language processing. The *syntactic validity label* (valid/invalid) indicates whether the conclusion logically follows from the premises according to formal syllogistic rules, independent of content truth. The *natural language understanding (NLU) label* (believable/unbelievable) indicates whether the conclusion is intuitively plausible given real-world knowledge, independent of logical structure.

The dataset comprises 76 valid syllogisms (47.5%) and 84 invalid syllogisms (52.5%). For believability, 38 instances (23.8%) have believable conclusions while 122 (76.2%) have unbelievable or abstract conclusions. This asymmetry reflects the inclusion of nonsense variants, which by design have semantically neutral conclusions.

**Belief Bias Categories** Belief bias is a well-documented phenomenon in human cognition whereby reasoners accept logically invalid conclusions that seem plausible, or reject valid conclusions that seem implausible—allowing the semantic content of conclusions to override evaluation of logical structure (Evans, Barston, and Pollard 1983; Klauer, Musch, and Naumer 2000; Pennycook et al. 2013).

Our dual annotation scheme enables formal quantification

of this effect by categorizing syllogisms based on alignment between logical validity and intuitive believability:

***Congruent instances*** (82 instances, 51.2%) are cases where logic and intuition align: valid-believable or invalid-unbelievable conclusions. These represent *"easy"* cases where correct logical judgment matches intuitive response.

***Incongruent instances*** (78 instances, 48.8%) are cases where logic and intuition conflict: valid-unbelievable or invalid-believable conclusions. These *"hard"* cases directly test whether models can override semantic plausibility with formal reasoning.

For example: *"All things with an engine need oil; Cars need oil; Therefore, cars have engines."* This conclusion is factually correct yet logically invalid (affirming the consequent fallacy). Such instances are particularly diagnostic, as accepting them indicates susceptibility to belief bias.

### 3.3 Prompting Schema

We implement four prompting strategies to evaluate models under varying levels of task specification and reasoning scaffolding: ***Zero Shot (ZS)*** and ***One-shot (OS)***, which utilize zero and one demonstration example respectively to test intrinsic capability; ***Few Shot (FS)***, which provides four balanced examples (2 valid, 2 invalid) including a belief bias trap to distinguish natural language plausibility from logical validity; and ***ZS Chain-of-Thought (ZS CoT)***, which encourages intermediate reasoning traces (Kojima et al. 2022). Critically, regardless of the context or scaffolding provided, all strategies request the same final response format: a single word *"correct"* or *"incorrect"* to ensure comparability across conditions.

Algorithm 1 presents our unified inference procedure that adapts its behavior based on the temperature parameter $\tau$. The algorithm accepts a syllogism $\mathcal{S}$ consisting of two premises $p_1, p_2$ and a conclusion $c$, a prompting strategy $\sigma$, and outputs a validity prediction $\hat{y}$ along with a confidence score $\rho$.

**Strategy Specifications** The procedure begins by constructing task-specific prompts through two subroutines. BUILDSYSTEMPROMPT($\sigma$) generates the system-level instruction that defines the reasoning task:

> "You are an expert in syllogistic reasoning. Your task is to determine whether the conclusion of a given syllogism follows from the premises. A syllogism is CORRECT if the conclusion follows from the premises. A syllogism is INCORRECT if the conclusion does not follow. ***[Strategy-specific addition.]*** Respond with exactly one word: 'correct' or 'incorrect'."

For ZS CoT, the system prompt appends *"Think through step by step"* before the response instruction; all other strategies use identical system prompts. BUILDUSERPROMPT($\mathcal{S}, \sigma$) constructs the user message by optionally including demonstration examples (1 for one-shot, 4 for FS), formatting the input syllogism with labeled premises and conclusion, and appending the query.

**Adaptive Stopping Strategy** When $\tau = 0$, the algorithm performs greedy deterministic decoding, querying the language model once, and returning the parsed prediction with full confidence ($\rho = 1.0$). For stochastic sampling ($\tau > 0$), we implement self-consistency (Chen et al. 2023) by generating up to $K_{\max} = 10$ independent samples. For each sample $k$, we query the model with temperature $\tau$ and parse the response to extract the validity label $\hat{y}_k$. We maintain counters $n_+$ and $n_-$ for valid and invalid predictions, respectively, using indicator functions $\mathbb{1}[\cdot]$.

To improve efficiency, we employ early stopping inspired by Holliday, Mandelkern, and Zhang (2024): if the first $\eta = 5$ samples are unanimous (i.e., $\min(n_+, n_-) = 0$ at $k = \eta$), we terminate sampling. This reduces API calls substantially when models exhibit high confidence. The final prediction $\hat{y}$ is determined by majority vote. Any ties by default maps to *"invalid"* as a conservative choice.

**Algorithm 1:** Temperature-Adaptive Syllogistic Reasoning

**Require:** Syllogism $\mathcal{S} = (p_1, p_2, c)$; Strategy $\sigma \in \{\text{ZS}, \text{OS}, \text{FS}, \text{ZSCoT}\}$; Temperature $\tau \in \{0.0, 0.5, 1.0\}$
**Ensure:** Prediction $\hat{y} \in \{\text{valid}, \text{invalid}\}$; Confidence $\rho \in [0, 1]$

**Parameters:** $K_{\max} = 10$, $\eta = 5$ {Max samples, early stopping threshold}

$\pi_{\text{sys}} \leftarrow$ BUILDSYSTEMPROMPT($\sigma$)
$\pi_{\text{user}} \leftarrow$ BUILDUSERPROMPT($\mathcal{S}, \sigma$)
**if** $\tau = 0$ **then**
  **return** PARSE(QUERY($\pi_{\text{sys}}, \pi_{\text{user}}, 0$)), 1.0
**end if**
$n_+ \leftarrow 0, n_- \leftarrow 0$
**for** $k = 1$ to $K_{\max}$ **do**
  $\hat{y}_k \leftarrow$ PARSE(QUERY($\pi_{\text{sys}}, \pi_{\text{user}}, \tau$))
  $n_+ \leftarrow n_+ + \mathbb{1}[\hat{y}_k = \text{valid}]$
  $n_- \leftarrow n_- + \mathbb{1}[\hat{y}_k = \text{invalid}]$
  **if** $k = \eta$ and $\min(n_+, n_-) = 0$ **then**
    **break** {Early stop if unanimous}
  **end if**
**end for**
$\hat{y} \leftarrow \begin{cases} \text{valid} & \text{if } n_+ > n_- \\ \text{invalid} & \text{otherwise} \end{cases}$ {Ties default to invalid}
$\rho \leftarrow \max(n_+, n_-)/(n_+ + n_-)$
**return** $\hat{y}, \rho$

### 3.4 Evaluation Methods

**Primary Metrics** We evaluate model responses using standard classification metrics: accuracy $(TP + TN)/N$, precision $TP/(TP + FP)$, recall $TP/(TP + FN)$, and F1 score as the harmonic mean of precision and recall. Accuracy serves as the primary metric given the near-balanced class distribution (47.5% valid, 52.5% invalid).

**Dual Evaluation Framework** Each model prediction is evaluated against both ground truths independently. For syntactic evaluation, the model response maps "correct" $\rightarrow$ *valid* and "incorrect" $\rightarrow$ *invalid*, compared against `ground_truth_syntax`. For NLU evaluation, it maps "correct" $\rightarrow$ *believable* and "incorrect" $\rightarrow$ *unbelievable*, compared against `ground_truth_NLU`. This dual evalua-

tion reveals whether models assess logical structure, natural language content, or some combination thereof.

**Belief Bias Effect** Classical belief bias research employed indices derived from raw endorsement rates (Evans, Barston, and Pollard 1983; Klauer, Musch, and Naumer 2000). However, these traditional indices have been criticized on psychometric grounds (Dube, Rotello, and Heit 2010; Heit and Rotello 2014): changes in proportions starting from different baseline values are not readily comparable, and empirical receiver operating characteristic (ROC) curves reveal curvilinear relationships that violate the linear assumptions of difference scores.

We adopt a direct accuracy-based approach aligned with recent studies (Trippas, Handley, and Verde 2014), quantifying belief bias as the accuracy differential between congruent and incongruent syllogisms:

$$\Delta_{\text{bias}} = \text{Acc}_{\text{congruent}} - \text{Acc}_{\text{incongruent}}$$

where $Acc_{congruent}$ is accuracy on valid-believable plus invalid-unbelievable instances (where logic and intuition align), and $Acc_{incongruent}$ is accuracy on valid-unbelievable plus invalid-believable instances (where they conflict).

This metric is appropriate for our setting because: (1) our LLM evaluations produce binary correct/incorrect judgments rather than confidence-rated responses, eliminating the ROC curvature concerns that motivated signal detection approaches (Dube, Rotello, and Heit 2010); (2) accuracy percentages are directly interpretable and comparable across all conditions, unlike endorsement-rate indices which suffer from baseline-dependency (Heit and Rotello 2014); (3) our within-subjects design compares each model against itself on congruent versus incongruent trials, isolating the belief bias effect while controlling for differences in overall reasoning ability. Positive $\Delta_{\text{bias}}$ indicates susceptibility to belief bias i.e., the model performs better when semantic content aligns with logical structure than when they conflict.

**Consistency Metric** We measure response consistency across content variants of logically equivalent syllogisms. Let $\mathcal{S}$ denote the set of 40 base natural syllogisms and $\hat{y}_{s,v}$ the model's prediction for syllogism $s$ under variant $v \in \{N, X, O, OX\}$. We define:

$$C_{\text{all}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{1}\left[\hat{y}_{s,N} = \hat{y}_{s,X} = \hat{y}_{s,O} = \hat{y}_{s,OX}\right] \quad (1)$$

$$C_{N \leftrightarrow X} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{1}\left[\hat{y}_{s,N} = \hat{y}_{s,X}\right] \quad (2)$$

$$C_{O \leftrightarrow OX} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbb{1}\left[\hat{y}_{s,O} = \hat{y}_{s,OX}\right] \quad (3)$$

where $C_{\text{all}}$ denotes overall consistency across all four variants. The pairwise metrics isolate specific invariance properties: $C_{N \leftrightarrow X}$ tests robustness to natural language content (meaningful vs. nonsense predicates), while $C_{O \leftrightarrow OX}$ tests robustness to premise order within matched content types.

# 4 Results

Our evaluation comprises 26,880 model-instance evaluations (14 models $\times$ 4 strategies $\times$ 3 temperatures $\times$ 160 syllogisms). We report syntactic accuracy as the primary metric, with supplementary analyses of dual-framework evaluation, belief bias, variant robustness, and response consistency.

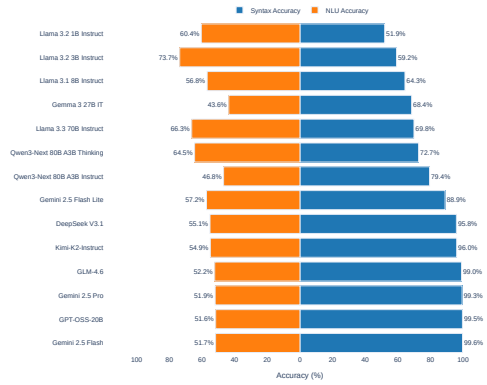


Figure 3: Syntactic validity (left) versus natural language understanding believability (right). The 25.50pp gap (syntax: 81.7%, NLU: 56.2%) demonstrates that models excel at formal logical reasoning while struggling with semantic plausibility judgments.

## 4.1 Overall Performance

Performance exhibits a bimodal distribution across the 14 evaluated models (Table 3). Six models achieve above 95% syntax accuracy, forming a distinct top-tier with robust syllogistic reasoning capability. Gemini 2.5 Flash attains near-perfect performance (99.6%), deviating from perfect accuracy in fewer than five instances per 1000. At the opposite extreme, five models score below 70%, with Llama 3.2 1B Instruct performing at 51.9%. The overall mean syntax accuracy is 81.7% ($SD$ = 17.1%), but the 47.7% gap between top and bottom performers demonstrates that syllogistic reasoning capability depends critically on architectural choices and training methods rather than raw model scale.

The pattern of precision, recall, and F1 scores reveals systematic biases. Qwen3-Next 80B A3B Thinking shows 99.2% precision but only 42.8% recall, indicating it labels most syllogisms as *"incorrect"* even when valid. Conversely, Gemma 3 27B IT exhibits 93.1% recall but only 61.0% precision, suggesting over-acceptance of conclusions. Top-tier models maintain balanced precision-recall profiles (both >97%), demonstrating genuine discriminative capability.

**Dual Evaluation Framework** We evaluated each prediction against both ground truths independently: *syntactic validity* and *NLU believability* (see §3.2). As shown in Figure 3 and Table 3 (final column), syntax accuracy (81.7%) substantially exceeds NLU accuracy (56.2%). Top-tier models show large syntax-NLU gaps: Gemini 2.5 Flash (47.9 pp), GPT-OSS-20B (47.9 pp), and Gemini 2.5 Pro (47.4 pp) excel at syntax but perform near chance on NLU evaluation. This pattern emerges because these models correctly judge logical validity independent of content believability.

| Model | Acc. | Prec. | Rec. | F1 | $C_{\text{all}}$ | $C_{N\leftrightarrow X}$ | $C_{O\leftrightarrow OX}$ | NLU Acc. |
|---|---|---|---|---|---|---|---|---|
| Gemini 2.5 Flash | 99.6 | 100.0 | 99.1 | 99.6 | 99.0 | 99.2 | 99.2 | 51.7 |
| GPT-OSS-20B | 99.5 | 100.0 | 99.0 | 99.5 | 96.5 | 97.1 | 98.1 | 51.6 |
| Gemini 2.5 Pro | 99.3 | 100.0 | 98.6 | 99.3 | 98.3 | 98.8 | 98.5 | 51.9 |
| GLM-4.6 | 99.0 | 100.0 | 97.8 | 98.9 | 95.8 | 96.5 | 97.5 | 52.2 |
| Kimi-K2-Instruct | 96.0 | 97.0 | 94.5 | 95.7 | 88.3 | 93.1 | 90.6 | 54.9 |
| DeepSeek V3.1 | 95.8 | 99.6 | 91.6 | 95.4 | 89.0 | 92.1 | 91.7 | 55.1 |
| Gemini 2.5 Flash Lite | 88.9 | 89.8 | 86.5 | 88.1 | 71.9 | 82.9 | 77.7 | 57.2 |
| Qwen3-Next 80B A3B Instruct | 79.4 | 73.3 | 88.9 | 80.4 | 69.2 | 81.0 | 76.5 | 46.8 |
| Qwen3-Next 80B A3B Thinking | 72.7 | 99.2 | 42.8 | 59.8 | 76.7 | 81.9 | 85.4 | 64.5 |
| Llama 3.3 70B Instruct | 69.8 | 82.1 | 46.7 | 59.5 | 66.2 | 81.0 | 78.3 | 66.3 |
| Gemma 3 27B IT | 68.4 | 61.0 | 93.1 | 73.7 | 69.0 | 82.5 | 86.0 | 43.6 |
| Llama 3.1 8B Instruct | 64.3 | 66.3 | 50.7 | 57.4 | 51.9 | 75.6 | 62.1 | 56.8 |
| Llama 3.2 3B Instruct | 59.2 | 88.1 | 16.2 | 27.4 | 75.0 | 92.1 | 81.7 | 73.7 |
| Llama 3.2 1B Instruct | 51.9 | 49.2 | 41.9 | 45.3 | 57.9 | 76.7 | 73.8 | 60.4 |

All metrics in %. Acc. = Syntax Accuracy, Prec. = Precision, Rec. = Recall.
Consistency metrics: $C_{\text{all}}$ (all 4 variants), $C_{N\leftrightarrow X}$ (normal $\leftrightarrow$ nonsense), $C_{O\leftrightarrow OX}$ (order-switched variants).

Table 3: Comprehensive model performance metrics aggregated across all 12 configurations (4 strategies $\times$ 3 temperatures). Syntax accuracy and NLU accuracy represent dual evaluation frameworks. Models grouped by performance tier.

Conversely, three models exhibit negative gaps: Llama 3.2 3B Instruct ($-14.5$ pp), Llama 3.2 1B Instruct ($-8.5$ pp), and Llama 3.3 70B Instruct ($+3.5$ pp shows minimal gap), suggesting that lower-tier models may rely more heavily on semantic plausibility heuristics.

## 4.2 Prompting Strategy Effects

Contrary to expectations, FS prompting yields the lowest mean accuracy (79.1%), while ZS achieves 82.7%. A paired $t$-test confirms that FS significantly underperforms ZS ($\Delta = -3.57$ pp, $t_{41} = 2.50$, $p = 0.0165$), with the effect surviving Holm-Bonferroni correction for three comparisons ($p_{\text{adj}} = 0.0495$, Cohen's $d = -0.39$). However, a Friedman test shows no significant overall strategy effect across all four strategies ($\chi^2 = 3.24$, $df = 3$, $p = 0.356$), and Wilcoxon signed-rank tests reveal the effect becomes marginally non-significant after correction ($p = 0.0195$, $p_{\text{adj}} = 0.0584$). Figure 1 illustrates the lack of systematic strategy effects across models.

To understand this pattern, we employed McNemar's test at the instance level ($N = 6720$ syllogism evaluations: 14 models $\times$ 3 temperatures $\times$ 160 syllogisms). We find highly significant error redistribution: ZS solves 786 instances that FS fails, while FS solves only 546 that ZS fails ($\chi^2 = 42.88$, $p < 0.0001$). The reconciliation is straightforward: FS prompting changes *which* problems are solved (McNemar test) and produces a consistent directional decline in mean accuracy ($t$-test), but the median effect is less robust (Wilcoxon test). Strategy effects appear model-specific rather than universal.

## 4.3 Temperature and Belief Bias Effects

Temperature ($\tau$) has negligible impact on accuracy when adaptive stopping is employed. A Friedman test confirms no significant temperature effect ($\chi^2 = 3.77$, $df = 2$, $p = 0.152$), with mean accuracy virtually identical across all $\tau$ settings. The adaptive majority-voting mechanism effectively normalizes stochastic variation.

| Model | Cong. | Incong. | $\Delta_{\text{bias}}$ |
|---|---|---|---|
| Llama 3.2 3B Instruct | 82.0 | 35.2 | +46.9 |
| Llama 3.3 70B Instruct | 85.3 | 53.6 | +31.6 |
| Qwen3-Next 80B A3B Thinking | 86.3 | 58.3 | +28.0 |
| Llama 3.2 1B Instruct | 62.0 | 41.2 | +20.8 |
| Llama 3.1 8B Instruct | 70.6 | 57.7 | +12.9 |
| Gemini 2.5 Flash Lite | 95.0 | 82.5 | +12.5 |
| DeepSeek V3.1 | 99.7 | 91.8 | +7.9 |
| Kimi-K2-Instruct | 99.6 | 92.1 | +7.5 |
| GLM-4.6 | 99.4 | 97.5 | +1.9 |
| Gemini 2.5 Pro | 100.0 | 98.6 | +1.4 |
| Gemini 2.5 Flash | 100.0 | 99.2 | +0.9 |
| GPT-OSS-20B | 99.2 | 98.4 | +0.8 |
| Qwen3-Next 80B A3B Instruct | 75.5 | 83.4 | −7.9 |
| Gemma 3 27B IT | 61.7 | 75.4 | −13.7 |

All values in %. Cong. = Congruent, Incong. = Incongruent.

Table 4: Belief bias analysis showing accuracy on congruent (logic matches intuition) versus incongruent (logic conflicts with intuition) syllogisms. Sorted by bias magnitude.

We observe robust evidence of belief bias across the majority of models (Figure 2, Table 4). Twelve of 14 models exhibit positive belief bias i.e., higher accuracy on congruent problems than on incongruent problems. The mean bias effect is $\Delta_{\text{bias}} = +10.81$ pp ($SD$ = 16.32), statistically significant by paired $t$-test ($t_{13} = 2.47$, $p = 0.0280$, Cohen's $d = 0.66$).

## 4.4 Consistency and Benchmark Correlations

The consistency metrics in Table 3 reveal that high-performing models maintain high consistency across content variants. The correlation between syntax accuracy and overall consistency is very strong (Pearson $r = 0.877$,

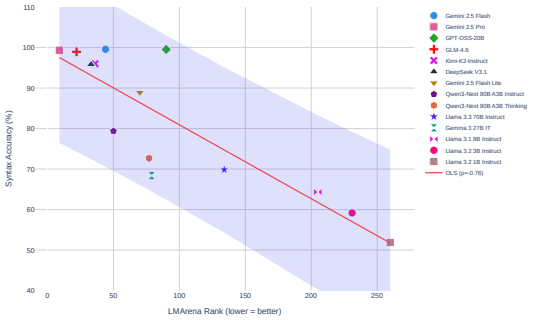


Figure 4: Correlation between syllogistic reasoning accuracy and LMArena rankings (Spearman $\rho = -0.825$, $p = 0.0010$, $N = 12$). Lower rank indicates better performance. The strong negative correlation suggests that instruction-following quality predicts formal reasoning capability.

$p < 0.0001$; Spearman $\rho = 0.890$, $p < 0.0001$), indicating that models achieving high accuracy are substantially more stable across variants.

To contextualize syllogistic reasoning within the broader LLM evaluation landscape, we computed correlations with LMArena human preference rankings (Chiang et al. 2024; Zheng et al. 2023, 2024). As shown in Figure 4, syllogistic reasoning shows a strong negative correlation with LMArena rank (Spearman $\rho = -0.825$, $p = 0.0010$, $N = 12$; lower rank indicates better performance). The negative correlation is the expected as models with higher reasoning accuracy achieve numerically lower (better) LMArena rankings. This suggests that models excelling at instruction following also excel at formal reasoning, likely because both require precise adherence to explicit rules.

### 4.5 Statistical Summary

Table 5 consolidates all key statistical findings. The FS underperformance survives Holm-Bonferroni correction ($p_{adj} = 0.0495$), while the McNemar test reveals significant error redistribution at the instance level. The reconciliation between significant $t$-test and marginally non-significant Wilcoxon test ($p_{raw} = 0.0195$, $p_{adj} = 0.0584$) reveals that FS produces a consistent mean decline but less robust median effect.

The correlation between syntax accuracy and belief bias magnitude shows a moderate negative relationship (Spearman $\rho = -0.565$, $p = 0.0353$). Since bias effect is defined as $Acc_{congruent} - Acc_{incongruent}$, this negative correlation indicates that higher performing models exhibit smaller bias magnitudes. It further provides evidence that higher reasoning ability reduces reliance on content based heuristics.

The very strong correlations between syntax accuracy and all three consistency metrics ($\rho = 0.890$, 0.846, and 0.837, all $p < 0.001$) confirm that models achieving high accuracy are substantially more stable across content and order variations. The moderate negative correlation between syntax and NLU accuracy (Spearman $\rho = -0.543$, $p = 0.0449$) indicates that models optimized for logical structure may diverge from intuitive believability judgments.

## 5 Discussion

In this study, we analyzed 40 instances of syllogism and their variations, resulting in a total of 160 data points tested against 14 different large language models. Our results demonstrate a striking pattern: top-tier models achieve near-perfect syntactic accuracy (99.6%) while performing at chance levels on natural language understanding (52%). This behavior, excelling at formal logic while struggling with semantic plausibility, contrasts sharply with human reasoning, where belief bias typically dominates logical analysis.

The majority of models exhibit significant belief bias, performing better when logic aligns with intuition (mean effect: +10.81 pp, $p = 0.028$). However, this bias decreases systematically with improved reasoning capability ($\rho = -0.565$, $p = 0.035$), suggesting that higher-performing models increasingly prioritize formal rules over semantic heuristics. Architectural and training choices prove more consequential than raw parameter count by substantial margins. Counter-intuitively, few-shot prompting degraded performance compared to zero-shot, suggesting demonstration examples may introduce noise in formal reasoning tasks. The strong correlation between instruction following quality (LMArena, $\rho = -0.825$) and reasoning accuracy indicates that precise rule adherence underlies both capabilities.

These findings suggest that most models exhibit a preference for symbolic reasoning and inferences rather than adhering to the natural language path of reasoning characteristic of human cognition. While this result may appear promising from a purely logical perspective, it raises important questions about the alignment between LLM reasoning and human cognitive processes. These models were trained on extensive natural language data, yet the top performers appear to function more like formal logic engines than human-like reasoners susceptible to known natural language biases.

## 6 Limitations

Our evaluation focuses primarily on categorical syllogisms, a narrow subset of logical reasoning that may not generalize to more complex structures with nested quantifiers or modal operators. The dual ground truth framework, while enabling systematic measurement, necessarily simplifies the dynamic interaction between logic and natural language that humans navigate simultaneously in real reasoning contexts.

The scope of our study includes only 14 models, representing a snapshot of the current LLM landscape but not exhaustive coverage of all available systems. Our prompting strategies, while covering major paradigms (zero-shot, one-shot, few-shot, chain-of-thought), constitute a limited exploration of the vast prompt engineering space. Additionally, our consistency metrics measure stability across content and order variations but do not assess robustness to adversarial perturbations or systematically manipulated distractors.

| Analysis | Test | Statistic | df | $p$-value | Effect | Result |
|---|---|---|---|---|---|---|
| *Main Effects* | | | | | | |
| Strategy effect (overall) | Friedman $\chi^2$ | 3.24 | 3 | 0.356 | — | No effect |
| ZS vs FS | Paired $t$ | 2.50 | 41 | $0.0165^{*}$ | $d = -0.39$ | Significant |
| ZS vs FS (Holm) | Paired $t$ | 2.50 | 41 | $0.0495^{*}$ | $d = -0.39$ | **Survives correction** |
| Temperature effect | Friedman $\chi^2$ | 3.77 | 2 | 0.152 | — | No effect |
| Belief bias (Cong. $>$ Incong.) | Paired $t$ | 2.47 | 13 | $0.0280^{*}$ | $d = 0.66$ | **Confirmed** |
| *McNemar Tests (Instance-level, $N = 6720$)* | | | | | | |
| ZS vs FS | McNemar $\chi^2$ | 42.88 | 1 | $<0.0001^{***}$ | 786 vs 546 | **Error redistribution** |
| ZS vs OS | McNemar $\chi^2$ | 1.70 | 1 | 0.192 | 317 vs 284 | No redistribution |
| ZS vs ZS CoT | McNemar $\chi^2$ | 0.26 | 1 | 0.612 | 389 vs 374 | No redistribution |
| *Key Correlations ($N = 14$ models)* | | | | | | |
| Syntax Acc. $\times$ Overall Consistency | Spearman $\rho$ | 0.890 | — | $<0.0001^{***}$ | Very strong | Positive |
| Syntax Acc. $\times$ $C_{N\leftrightarrow X}$ | Spearman $\rho$ | 0.846 | — | $0.0001^{***}$ | Very strong | Positive |
| Syntax Acc. $\times$ $C_{O\leftrightarrow OX}$ | Spearman $\rho$ | 0.837 | — | $0.0002^{***}$ | Very strong | Positive |
| Syntax Prec. $\times$ Syntax Rec. | Spearman $\rho$ | 0.691 | — | $0.0062^{**}$ | Strong | Positive |
| Syntax Acc. $\times$ NLU Acc. | Spearman $\rho$ | $-0.543$ | — | $0.0449^{*}$ | Moderate | **Negative** |
| Syntax Acc. $\times$ Bias Effect | Spearman $\rho$ | $-0.565$ | — | $0.0353^{*}$ | Moderate | **Negative** |
| *Benchmark Correlation* | | | | | | |
| LMArena rank (lower = better) | Spearman $\rho$ | $-0.825$ | — | $0.0010^{***}$ | Very strong | **Predicts reasoning** |

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Holm-Bonferroni correction applied to strategy comparisons.
McNemar instances: "786 vs 546" = ZS correct & FS wrong vs FS correct & ZS wrong.
Bias correlation: Negative $\rho$ means higher accuracy correlates with smaller bias magnitude (closer to zero).

Table 5: Comprehensive statistical summary of all hypothesis tests and correlations for 14 models. Strategy comparisons use Holm-Bonferroni correction. McNemar test operates at instance-level (6,720 syllogism evaluations per comparison).

The belief bias metric, while grounded in cognitive psychology literature, captures only one dimension of the complex relationship between real world beliefs and logical reasoning. Future work should incorporate additional measures such as response time analysis, confidence calibration, and fine-grained error taxonomies to provide a more comprehensive understanding of LLM reasoning processes.

## 7 Future Work

Several promising directions emerge from this work. Extending evaluation to richer logical systems such as modal logics, transitive closure logics, to test whether the observed patterns generalize beyond categorical syllogisms. Particular interest lies in logical systems with simple formal syntax but complex natural language semantics, which would further stress the formal logic-natural language divide that we observed.

Complementing these empirical extensions, mechanistic interpretability studies could reveal whether models learn explicit logical rules, statistical approximations, or hybrid representations. This would clarify the computational basis of the near-perfect syntactic performance we documented in top-tier models. Related to this, the causal relationship between reasoning capability and bias resistance remains an open question: does logical training reduce bias, or does reduced bias enable better reasoning? Controlled fine-tuning experiments could disentangle these possibilities.

Our finding that few-shot prompting degraded performance challenges conventional wisdom and warrants systematic exploration of when and why demonstration examples help versus hinder reasoning. Such investigation would inform more effective prompting strategies for logical reasoning tasks.

More broadly, our results raise a fundamental tension i.e., are we building human like reasoners or formal logic engines? This question has implications not only for model development but also for appropriate deployment contexts and expectations for LLM behavior in reasoning-intensive applications. We intend to continue this line of inquiry across other logical reasoning tasks to better understand the trajectory of the cognitive capabilities of LLM.

## Acknowledgements

We thank the Indo-French Centre for the Promotion of Advanced Research (IFCPAR/CEFIPRA) for their support. This work was supported through project number CSRP-6702-2.

## References

Borazjanizadeh, N.; and Piantadosi, S. T. 2024. Reliable Reasoning Beyond Natural Language. *arXiv e-prints*, arXiv–2407.

Braüner, T.; Ghosh, A.; and Ghosh, S. 2025. Understanding responses of people with ASD in diverse reasoning tasks: A formal study. *Cognitive Processing*, 26(1): 201–218.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; and Horvitz, E. 2023. Ece 1251 Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general 1252 intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 1253.

Chen, X.; Aksitov, R.; Alon, U.; Ren, J.; Xiao, K.; Yin, P.; Prakash, S.; Sutton, C.; Wang, X.; and Zhou, D. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132.

Copi, I. M.; Cohen, C.; and McMahon, K. 2016. *Introduction to logic*. Routledge.

Dube, C.; Rotello, C. M.; and Heit, E. 2010. Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, 117(3): 831–863.

Eisape, T.; Tessler, M.; Dasgupta, I.; Sha, F.; Steenkiste, S.; and Linzen, T. 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)*, 8425–8444.

Evans, J. S. B.; Barston, J. L.; and Pollard, P. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3): 295–306.

Heit, E.; and Rotello, C. M. 2014. Traditional difference-score analyses of reasoning are flawed. *Cognition*, 131(1): 75–91.

Holliday, W.; Mandelkern, M.; and Zhang, C. 2024. Conditional and Modal Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3800–3821.

Kim, G.; Valentino, M.; and Freitas, A. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, 10074–10095.

Klauer, K. C.; Musch, J.; and Naumer, B. 2000. On belief bias in syllogistic reasoning. *Psychological review*, 107(4): 852.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lewton, M. 2016. *The relationship between autism and psychosis traits and reasoning style*. Ph.D. thesis, University of Bath.

Pennycook, G.; Cheyne, J. A.; Koehler, D. J.; and Fugelsang, J. A. 2013. Belief bias during reasoning among religious believers and skeptics. *Psychonomic Bulletin & Review*, 20(4): 806–811.

Sambrotta, M. 2025. LLMs and the Logical Space of Reasons. *Minds and Machines*, 35(4): 1–23.

Smith, R. 2022. Aristotle's Logic. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Smith, R.; et al. 1989. *Prior analytics*. Hackett Publishing.

Solcz, S. 2008. *The Role of Working Memory in Deductive Reasoning: A Dual Task and Individual Differences Approach*. Ph.D. thesis, University of Waterloo.

Trippas, D.; Handley, S. J.; and Verde, M. F. 2014. Fluency and belief bias in deductive reasoning: New indices for old effects. *Frontiers in Psychology*, 5: 631.

Zhang, K.; Yu, W.; Sun, Z.; and Xu, J. 2025. Syler: A framework for explicit syllogistic legal reasoning in large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 4117–4127.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Li, T.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Li, Z.; Lin, Z.; Xing, E.; Gonzalez, J. E.; Stoica, I.; and Zhang, H. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *The Twelfth International Conference on Learning Representations*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zong, S.; and Lin, J. 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)*, 230–239.