Evaluating LLMs on Syllogistic Reasoning: A Human–Model Accuracy Comparison under Premise Order Effects

Anonymous submission

Abstract

Large Language Models (LLMs) have achieved remarkable success on many natural language tasks; however, their performance in logical reasoning is still unsatisfactory. The aim of this study is to evaluate LLMs' sensitivity to premise order in logical reasoning within the classic paradigm of categorical syllogisms, and to benchmark their accuracy against a human baseline. We constructed a test set of 64 natural language syllogisms with a dual-order design. Human participants (N=1317) randomly completed 32 items (16 forward-order, 16 reverse-order), while 12 LLMs completed all 64 items. Using accuracy as the sole metric, we defined the order effect as $\Delta acc = Acc_{forward} - Acc_{reverse}$ and conducted statistical analyses at the overall, per-figure, and per-form levels across 32 logical forms. The results show that while the human group exhibits no overall order effect, LLMs as a whole display a weak and non-systematic effect with inconsistent directions across different models. Furthermore, the models proved to be more fragile on logical forms that were challenging for both humans and machines. The human–LLMs correlation of Δacc across the 32 formats is nearly zero, and the directional agreement rate is not significantly higher than chance. Our work provides a new conceptual framework and an empirical benchmark for investigating intrinsic limitations of LLM reasoning.

Introduction

Large language models (LLMs) have a systematic sensitivity to the ordering of input information, which has been shown to be a core challenge to their reasoning robustness. This vulnerability commonly manifests as position bias, whereby a model's decisions are influenced by the order in which options or answers appear in the prompt. Such bias has been widely documented across multiple task paradigms, including candidate ranking in recommender systems (Hou et al. 2024), option ordering in multiple-choice questions (Pezeshkpour and Hruschka 2023), and answer ordering in "LLM-as-a-judge" tasks (Wang et al. 2023; Zheng et al. 2023; Wu and Aji 2023).

Regarding the source of this bias, a common hypothesis is that LLMs inherit human cognitive biases (Sumita, Takeuchi, and Kashima 2025), for example reproducing the human primacy effect in choice tasks (Wang et al. 2024). However, this inheritance hypothesis is not conclusive. Some work argues that LLMs' apparent reasoning may simply reflect a "Clever Hans effect," relying on structural heuristics such as option order rather than genuine reasoning (Ranaldi and Zanzotto 2024). Other studies propose that the root of LLM bias may not be positional order perse, but a prior token bias toward specific option ID characters (e.g., "A") (Zheng et al. 2024).

Existing work has amply demonstrated LLMs' irrational vulnerability to option order and sequence perception. However, these studies largely focus on the surface task of option selection and mostly assume that the bias stems from position or from inherited human biases. At present, there is little research on a more fundamental order effect—namely, sensitivity in purely logical reasoning to reversing two logically equivalent premises (P1, P2 vs. P2, P1). For instance, existing work has focused on whether a model's answer to "Which city is the capital of France?" is affected by the order of options, such as "(A) Paris, (B) London" versus "(A) London, (B) Paris". Our study, in contrast, investigates a more fundamental logical operation: does a model's conclusion from the premises "All men are mortal; Socrates is a man" differ from the conclusion derived from "Socrates is a man; All men are mortal"? The former tests surface-level choice-making, while the latter probes the core process of logical integration. To fill this gap, this study introduces a "pure premise-order effect" paradigm, employing syllogistic reasoning to probe the internal mechanisms of cognitive integration in LLMs. Crucially, and for the first time on a large scale, we conduct a direct human-model comparison to test whether the observed bias patterns in LLMs genuinely originate from human cognitive heuristics.

The syllogism, as the fundamental framework for deductive logic, provides a formal structure whose normativity is a priori, guiding both human cognition and language generation. For any ideal rule-based reasoning system, logically equivalent inputs should yield invariant outputs. Syllogistic laws have long been considered normative principles governing the operation of the rational mind (Russell and Norvig 2020). Based on this principle, the syllogism provides a decontextualized and atemporal, purely formal testing environment. It allows for the examination of not only the correctness of the inferred conclusion but also the invariance of properties under logically equivalent transformations, making it an ideal framework for assessing the "logical consistency" of any reasoning agent, which we define

here as the invariance of outputs under logically equivalent transformations of inputs.

To address the aforementioned questions from an empirical perspective, this study adopts the "order effect" paradigm from syllogistic reasoning to systematically examine the impact of premise order (forward/reverse)—a logically equivalent transformation—on reasoning accuracy. We construct a strictly controlled set of natural-language syllogisms, forming paired items solely by permuting the positions of the major and minor premises, in order to test whether this non-logical factor affects the judgment outcome. To establish a robust empirical baseline, we include a large human cohort (N=1317) as a baseline and compare 12 mainstream LLMs under the same test set and evaluation protocol.

The core contribution of this study is to propose a logically symmetric and formally complete syllogistic evaluation framework. This framework comprises 64 items in a dual-order paired design, ensuring comprehensive coverage of all figures and moods. The framework uses accuracy as the primary metric, while also introducing "order invariance" (Δacc) to characterize stability under logically equivalent inputs. Under the same test set and protocol, we provide a systematic comparison between the human baseline and 12 LLMs: (1) the human group exhibits no overall order effect, which aligns with the principles of categorical logic; (2) LLMs as a whole display a weak and non-systematic effect with inconsistent directions; and (3) models are more prone to instability on logical forms that are challenging for both humans and machines, such as combinations of negative and existential premises or weak moods.

Background

The core structure of the syllogism was first systematically articulated by the ancient Greek philosopher Aristotle (384–322 BC) in the Organon. He defined the syllogism as "a form of reasoning in which the conclusion necessarily follows from the premises", and explicitly distinguished the major premise, the minor premise, and the conclusion (Aristotle, Prior Analytics 24b18–20).

In the Prior Analytics, Aristotle formulated syllogisms in natural language (e.g., "All A are B; all B are C; therefore all A are C"), and termed the premise containing the major term (the predicate of the conclusion) the major premise, the premise containing the minor term (the subject of the conclusion) the minor premise, with the conclusion being the connection between the major and minor terms. In the axiomatic system of classical categorical logic, the validity of a syllogism is strictly determined by the formal structure defined by its figure and mood. The order of a syllogism refers to the two different arrangements of the major and minor premises.

Aristotle posited that the acquisition of knowledge begins with grasping universal principles (major premise), proceeds through empirical induction to subsume particular instances under universal categories (minor premise), and culminates in definitive conclusions. For example:

- Major Premise (Universal Principle): All bileless creatures are long-lived.
- Minor Premise (Particular Instance): Humans and horses are bileless creatures.
- Conclusion (Necessary Judgment): Humans and horses are long-lived.

Medieval logicians (such as Boethius and Thomas Aquinas) translated Aristotle's Greek logical works into Latin and standardized the structure of syllogisms. For instance, they fixed the terminology and order as *Major Premissa* (major premise), *Minor Premissa*(minor premise), and *Conclusio*(conclusion), forming the standard presentation *major premise + minor premise* $\rightarrow$ *conclusion*.They further developed the Four Figures of Syllogism (classified by the position of the middle term) and identified the 24 Valid Moods.

From the 18th century onward, logic textbooks (e.g., the appendix to Kant's *Critique of Pure Reason*) simplified syllogism into the cognitive model: "Major Premise (Principle) + Minor Premise (Application) $\rightarrow$ Conclusion (Result)". Thus, the formal structure of the Four Figures and 24 Moods became stabilized and widely recognized as the canonical logical framework.

The order of a syllogism refers to the two different sequences of the major and minor premises. In the canonical form of a syllogism, the major premise precedes the minor premise. Considering the order of the two premises, syllogism can have the following eight forms of premises:

Figure	**First**		**Second**		**Third**		**Fourth**	
	P1	P2	P1	P2	P1	P2	P1	P2
Case 1	M-P	S-M	P-M	S-M	P-M	M-S	P-M	M-S
Case 2	S-M	M-P	S-M	P-M	M-S	P-M	M-S	P-M

The traditional logic recognized four forms of propositions: **A** (All S are P), **E** (No S are P), **I** (Some S are P), and **O** (Some S are not P).(Abstractly, the subject category is named S and the predicate category is named P.) The letters **A** and **I**, from the Latin word *affirmo*, stand for 'affirmative' propositions. The letters **E** and **O**, from *nego*, stand for 'negative' propositions.

It is generally held that these two arrangements make no difference to the syllogism's validity, as this conforms to the commutative law of logical conjunction ($A \wedge B \Leftrightarrow B \wedge A$), which ensures that the logical consequence of a set of premises does not change with the permutation of its elements. Therefore, under the same form, permuting the premises does not alter the logical consequence.

Dataset Statistics

Experimental Materials

The experiment contains a dataset of 64 natural language syllogisms, expertly constructed to systematically evaluate logical reasoning. This dataset provides comprehensive coverage of 32 unique syllogistic forms, encompassing all four figures and a curated selection of moods designed to span a range of difficulties.

To investigate the pure premise-order effect, each logical form was instantiated in two versions that are strictly logically equivalent and differ only in the order of premises. To minimize potential confounds from language comprehension, background knowledge, and corpus patterns—and to ensure that task difficulty focuses purely on logical reasoning—the dataset was compiled under stringent principles of linguistic simplicity and naturalness, and all premises and conclusions were written as contingently true statements.

- Example of a forward-order item:

Given the following two premises:All flowers planted by Uncle Wang are rare species.All plants in the main garden were planted by Uncle Wang.Can it be deduced that: All plants in the main garden are rare species?Yes ( ), No ( ), Uncertain ( )

- Example of a reverse-order item:

Given the following two premises: All plants in the main garden were planted by Uncle Wang.All flowers planted by Uncle Wang are rare species.Can it be deduced that: All plants in the large garden are rare species?Yes ( ), No ( ), Uncertain ( )

Participants and Procedure

Human Participants We recruited a total of 1,385 human participants, each of whom completed one 32-item syllogism test (16 forward-order and 16 reverse-order items). To ensure that the two questionnaires had a high degree of equivalence in terms of difficulty, we conducted an independent samples t-test on the scores of the participant groups who completed the different versions. The results of the test showed no statistically significant difference between the mean scores of the two groups ($t(1383) = 1.44, p = 0.149$). This indicates that *QNR 1* and *QNR 2* can be considered equivalent in overall difficulty, thus justifying the pooling of the two human datasets for subsequent analysis.

QNR	N	M	SD	t-Statistic (df)	p-Value
1	700	17.1	5.25	1.4439	0.1490
2	685	16.7	4.93		

Table 1: Analysis of Questionnaire Equivalence.

This study used only accuracy (or total score) as the sole evaluation metric and did not collect or analyze reaction times, confidence ratings, or explanatory texts. To ensure the robustness of the analysis, we performed outlier removal based on the distribution of total scores (the number of correct responses out of 32 items per participant). Based on the **mean** (M) and **standard deviation** (SD) of the entire sample, we retained participants whose total scores fell within the range of $M \pm 2SD$. A total of 68 outliers (4.91% of the initial sample) were excluded, resulting in a final effective sample of 1,317 participants for subsequent statistical analyses and the human–machine comparison.

Large Language Models This study selected 12 representative LLMs as test subjects. The selection of models was designed to cover a wide range of current mainstream closed-source and open-source models. It encompasses different research and development institutions (e.g., OpenAI, Google, Meta, Anthropic, and leading Chinese AI companies), various technical architectures (dense models vs. Mixture-of-Experts, or MoE), and diverse parameter scales, thereby ensuring the breadth and representativeness of our evaluation. A detailed list of the models and their characteristics is provided in Table S1

The rationale for this selection (omitted from Table S1 for brevity) was to ensure a diverse and representative sample. This included: (1) Flagship models from major international and domestic organizations (e.g., OpenAI, Google, Anthropic, Zhipu AI) to establish performance baselines. (2) A mix of architectures, such as dense models and Mixture-of-Experts (MoE). (3) Models spanning a wide range of parameter scales, from lightweight (˜7B) to ultra-large-scale (˜235B). (4) Both closed-source APIs and prominent open-source models. (5) Models with specific optimizations for tasks like long-context, multimodal, or multilingual reasoning.

All models were tested via their official APIs or standard interfaces, with each model being evaluated on each item only once. To minimize the randomness of the output, the decoding temperature for all models was set to 0 or the lowest possible value approaching 0. We provided the models with the complete problem, including the premises and the conclusion, and explicitly instructed them to "output only one word: Yes, No, or Uncertain." Returned outputs were scored using strict keyword matching: if a response begins with one of the three specified tokens, it was treated as a valid answer and evaluate for correctness; any other form of output was treated as invalid and counted as incorrect.

Experiments

This section details the empirical results of the experiment. First, the data from the 1317 human participants are analyzed. No significant overall premise order effect was detected in human participants; however, statistically significant and bidirectional local effects were observed at the perform level. Second, the performance of the 12 LLMs is evaluated. The results show that the LLMs' response to the overall order effect, while also not statistically significant, is descriptively stronger and more directionally consistent compared to the human baseline. In stark contrast to humans, however, LLMs show no significant local effects at the perform level. Finally, the third part provides a direct human-LLM comparison of the two patterns. This comparison confirms a fundamental "decoupling" between the two groups: the premise order effect patterns of humans and LLMs are statistically uncorrelated. This reveals that the models' intrinsic mechanism for processing premise order is statistically independent from that of humans, and their reasoning biases are not the same.

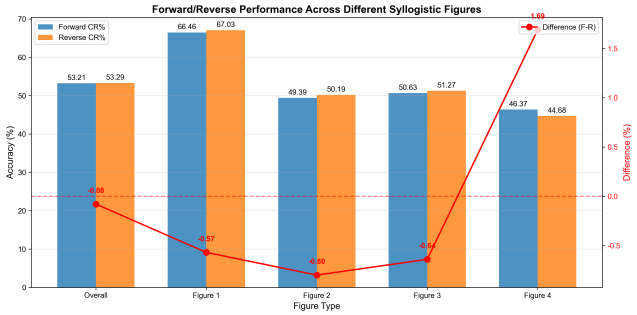

Figure 1: Performance of Human Participants on Different Figures.

Human

First, we analyzed the performance of the 1317 human participants on the syllogistic reasoning task. The overall mean accuracy for the participant group was 53.25% ($M = 17.04$, $SD = 5.09$). At the aggregate level, we examined the effect of premise order on human reasoning accuracy. As shown in Figure 1, the accuracies for forward and reverse orders were highly similar both in overall averages and when disaggregated by the four figures.

A paired-samples t-test on the total scores also confirmed this observation, revealing no statistically significant difference between the mean scores for the forward-order ($M = 8.51$, $SD = 2.39$) and reverse-order ($M = 8.53$, $SD = 2.45$) conditions ($t(1316) = -0.21$, $p = .834$). This result indicates that, at the aggregate level, the human participants did not exhibit an overall order effect.

However, item-wise tests of the order effect across the 32 logical forms (Δacc = Forward Accuracy − Reverse Accuracy) revealed significant order differences for a minority of forms (Chi-square tests, $p < 0.05$).

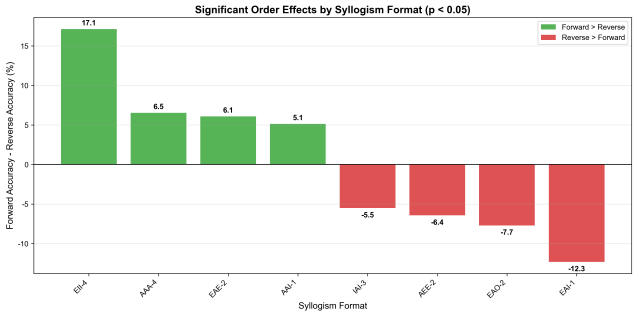

Figure 2: Forms with Significant Premise Order Effects in Human Participants.

As shown in Figure 2, the influence of premise order reached a statistically significant level for 8 specific logical forms. These significant effects demonstrated strong, opposing directional characteristics. Among them, 4 forms showed a significant forward-order advantage (i.e., the forward-order task was easier), such as in the EII-4, where the forward-order accuracy was 17.13 percentage points higher than the reverse-order accuracy ($p < 0.001$). The other 4 forms showed a significant reverse-order advantage (i.e., the reverse-order task were easier); for instance, in EAI-1, where reverse accuracy exceeded forward accuracy by 12.32 percentage points ($p < 0.001$).

To provide a more comprehensive view of this phenomenon, Figure 3 presents the distribution of order effects across all 32 logical forms. As shown, Δacc is broadly distributed on both sides of zero, with both forward and reverse advantages coexisting and varying in magnitude (see Figure 3).

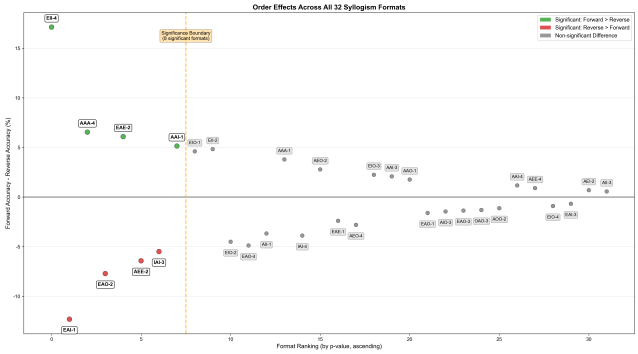

Figure 3: Full Spectrum of Premise Order Effects across All 32 Forms for Human Participants.

This indicates that the overall "no order effect" stems from the mutual cancellation of bidirectional differences at the per-form level, rather than from a globally consistent insensitivity to order. For complete per-form values and statistical tests, see Table S2- S4 .

Large Language Models

First, we evaluated the overall reasoning accuracy of each model on the full set of 64 items. As shown in Table 2, the overall accuracies of the LLMs reveal a clear stratification of capabilities. Llama-4-Scout-17B-16E-Instruct ranked first with an accuracy as high as 96.88%, indicating strong performance in logical reasoning. It was closely followed by GPT-4o (93.75%) and Qwen3-235B (92.19%), which constitute the top tier of performance. Notably, several models, including DeepSeek-V3, Ling-max-1.5, and even the small-sized Qwen3-8B, clustered at an accuracy level of 89.06%. Compared to the human baseline mean accuracy of 53.25%. This suggests that, on purely formal logical reasoning tasks, current mainstream LLMs achieve accuracy substantially exceeding that of typical non-expert populations.

Model Name	Acc_forward (%)	Acc_reverse (%)	Overall Acc (%)	Δacc (%)
Llama-4-Scout-17B-16E-Instruct	100.00	93.75	96.88	6.25
GPT-4o	96.88	90.62	93.75	6.25
Qwen3-235B-A22B-Instruct	87.50	96.88	92.19	−9.38
DeepSeek-V3	90.62	87.50	89.06	3.12
Qwen3-8B	90.62	87.50	89.06	3.12
Claude-3.7-Sonnet	84.38	90.62	87.50	−6.25
Seed-OSS-36B-Instruct	87.50	81.25	84.38	6.25
Gemini-2.5-Pro	84.38	81.25	82.81	3.12
DeepSeek-R1	75.00	75.00	75.00	0.00
GLM-4.5	75.00	62.50	68.75	12.50
gpt-oss-120b	62.50	68.75	65.62	−6.25
Moonshot-v1-8k	62.50	65.62	64.06	−3.12
Human Baseline	53.18	53.31	53.25	-0.13

Table 2: Overview of Reasoning Performance for the 12 Large Language Models.

Under the same test set and evaluation protocol, LLMs as a whole demonstrated a very small difference in mean accuracy between the forward-order ($M = 82.81\%$) and reverse-order ($M = 80.86\%$) tasks ($\Delta = +1.95\%$), exhibiting a weak "forward-order advantage." Disaggregated by figure, the overall LLM accuracies for forward vs. reverse order are: Figure 1, 96.09% vs. 94.53%; Figure 2, 82.81% vs. 82.03%; Figure 3, 87.50% vs. 85.94%; Figure 4, 64.84% vs. 60.94%. All four figures exhibit a modest forward-order advantage. As shown in Figure 4

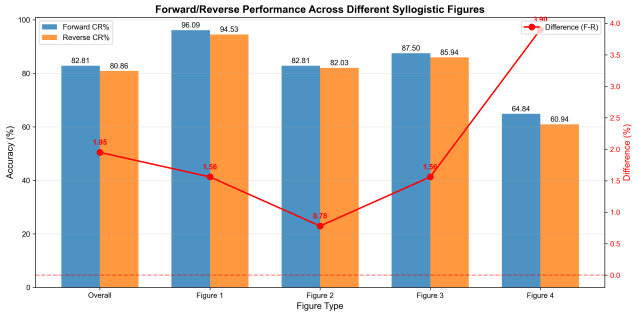

Figure 4: Performance of LLMs on Different Figures.

Aggregating Δacc (Forward − Reverse) across the 32 logical forms into an overview plot shows that most forms cluster near zero, with small deviations on both sides, typically within ±12.5 percentage points. In conjunction with significance tests (total forms: 32; $p < 0.05$: 0; $p < 0.20$: 4), the overall order sensitivity of LLMs is weak at the form level and directionally inconsistent (see Figure 6). Consistent with the figure–wise results in Figure 4, the LLMs' weak forward-order advantage is primarily an average–level phenomenon; at the per–form level, order differences are mostly small fluctuations, with a few larger positive/negative values that do not reach statistical significance. Mixed-effects model summaries (order-effect coefficients, 95% CIs,

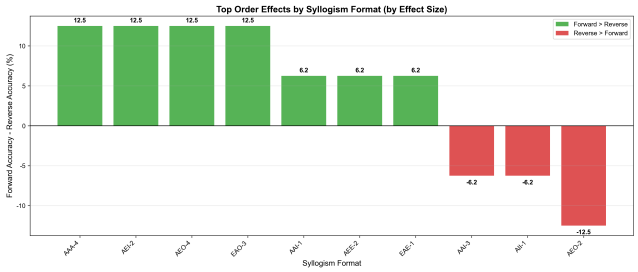

Figure 5: Top Premise Order Effects in LLMs by Effect Size.

p-values) are listed in Table S7 ; pooled odds ratios for models are summarized in Table S5.

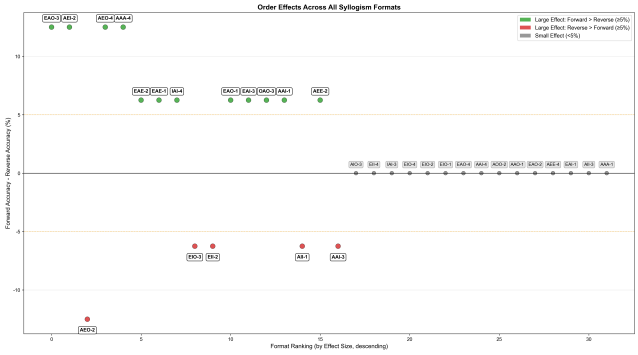

Figure 6: Full Spectrum of Premise Order Effects across All 32 Forms for LLMs.

Human–LLM Comparison

Having analyzed human and LLM performance separately, this section provides a direct comparison of their respective order effect patterns. We begin by comparing the overall trend of the order effect across the 32 forms for both groups of participants at a macroscopic level(Figure 7). As noted above, the mean order effect in humans (Δacc = Forward Accuracy − Reverse Accuracy) is near zero ($M = -0.07\%$, 95% CI $[-1.97\%, 1.83\%]$), whereas the aggregate mean order effect for LLMs shows a slight positive trend ($M = +1.95\%$, 95% CI $[-0.22\%, 4.13\%]$), reflecting a non-significant "weak forward-order advantage." Although both means appear close to zero, a paired-samples t-test indicates that the difference between the two patterns is not statistically significant ($t = -1.39$, $p = .174$).

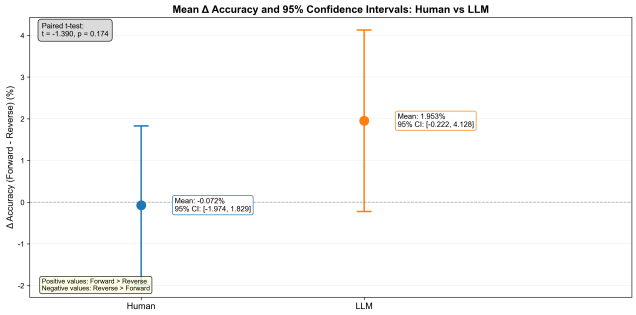

Figure 7: Contrasting the Mean Premise Order Effect between Humans and LLMs.

The proximity of aggregate means does not imply a correspondence between the two patterns. To investigate whether human order preferences can predict LLM preferences, we conducted a correlation analysis on the magnitude of the order effect (Δacc) for both groups across the 32 logical forms. The scatter plot (Figure 8) shows that the data points, each representing a syllogistic form, do not cluster around the $y = x$ diagonal but instead exhibit a diffuse and irregular distribution. Correlation analyses statistically corroborate this visual observation: there is no significant linear association between human and LLM order-effect magnitudes (Pearson's $r = 0.052$, $p = 0.775$; Spearman's $\rho = 0.023$, $p = 0.899$) (see Table S8 for underlying data).

The ten "human-significant" forms ($p < 0.05$) marked with red stars in the plot are distributed at the extremes of the x-axis (human Δacc), spanning from strongly negative effects (EAI-1: -13.15%, EAO-2: -8.17%) to strongly positive ones (EII-4: $+17.07\%$, AAA-4: $+6.84\%$). However, these forms exhibit two distinct patterns on the y-axis (LLM Δacc). First, for the majority of these forms (6 out of 10: EII-4, EAI-1, EAO-2, EAO-4, EIO-2, AAA-1), the LLM effect is zero, indicating no order sensitivity despite strong human effects. Second, for the remaining four forms, LLMs do show an order effect, but with varying alignment to human patterns. Three forms (AAA-4: $+12.50\%$, AAI-1: $+6.25\%$, EAE-2: $+6.25\%$) exhibit concordant forward-order advantages in both populations, suggesting partial alignment. However, AEE-2 presents a striking dissociation: humans show a reverse-order advantage (-5.90%) while LLMs show a forward-order advantage ($+6.25\%$), demonstrating directly opposing preferences. At the global level, direction concordance is weak (McNemar's test, $p = 0.7539$; see Table S8 for the contingency table), and a complete per-form list of directional labels and concordance flags is provided in Table S9.

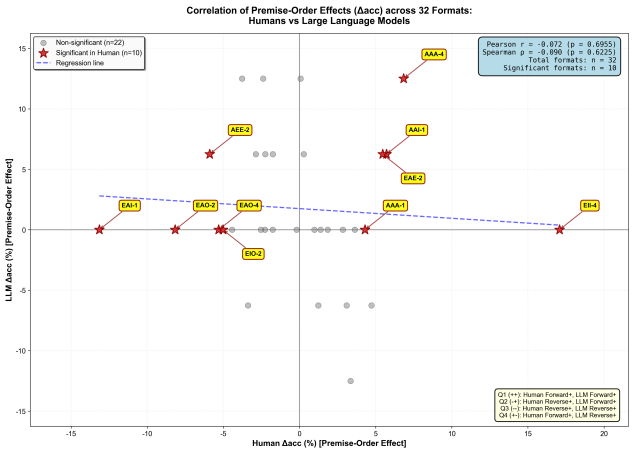

Figure 8: Correlation of Premise-Order Effects (Δacc) across 32 Formats: Humans vs LLMs.

Discussion

This study systematically examines the logical reasoning capabilities of LLMs through a human–machine comparison experiment grounded in syllogistic premise-order effects. The core finding is that, although mainstream LLMs achieve accuracy substantially exceeding that of typical non–experts on formal reasoning tasks, their behavior under logically equivalent transformations exhibits a form of inconsistency that is markedly different from that observed in humans.

Divergence in Reasoning Patterns: Human vs. LLM

This study precisely distinguishes two different patterns of reasoning error, through human–machine comparison. First, the human reasoning pattern can be characterized as "globally compliant, locally systematic deviation." In classical categorical logic, premise permutation does not affect validity, and our data show that the human group indeed exhibits no overall order effect in this paradigm ($\Delta\text{acc} \approx 0$, t not significant). This suggests that when the task is strictly confined to making logical judgments based solely on the premises, the human group as a whole adheres to the principle of formal invariance in natural language syllogisms.

However, at the per–form level we observe strong "local order effects": for specific forms like EII–4 and EAI–1, changes in premise order significantly impact reasoning difficulty, leading to predictable, systematic biases in oppos-

ing directions. This indicates that human reasoning does not have a preference for *per se*; rather, it is the interaction between their internal cognitive processing paths (e.g., strategies for handling negation and existential quantifiers) and specific logical structures that creates an asymmetric cognitive load. In contrast, LLMs as a whole exhibit a weak and non–significant forward advantage ($\Delta\text{acc} = +1.95\%$), with no statistically significant effect detected for any single form. However, this apparent robustness conceals an underlying fragility. The direction and magnitude of the order effect vary across different models (e.g., a reverse-order advantage for Qwen3-235B vs. a forward-order one for GPT-4o). This suggests that "order sensitivity" is a dimension independent of a model's overall accuracy and is likely influenced by specific model architectures (such as expert activation paths in MoE models) or implementation strategies. This behavior is more akin to an unpredictable "computational accident" dependent on the specific input token sequence, rather than a regular cognitive bias. Finally, we also observed "common difficulties" shared by both humans and machines. On some logically more complex forms, such as the EII mood involving a combination of negation and existence, and the weak AAA–4 mood which is invalid under classical logic due to existential import issues, both groups showed low accuracy. This suggests that precise mapping from natural language to categorical logic is intrinsically challenging, and that the inherent complexity of certain logical forms (e.g., term distribution and existential import) constitutes a common bottleneck for humans and current LLMs.

Mechanistic Analysis

The data from the human sample is characterized by a "null overall order effect, with significant local effects." This suggests that while the validity of a syllogism is independent of premise order, the way natural language presents the premises interacts with human processing paths to induce biases in specific structures. For instance, the forward-order advantage in the EII–4 form may stem from a linear strategy of "processing the universal negative premise first, then the existential quantifier," which could effectively reduce the burden of constructing mental models. Conversely, the reverse-order advantage in EAI–1 might be because "reading the affirmative premise first" reduces the cognitive interference caused by negative information. These are all cognitive strategies that the human brain has evolved under limited resources, aimed at "taking shortcuts."

In contrast, the data from the model sample reflects the characteristics of a "weak order effect with inconsistent directions." For Transformer–based models, changing premise order fundamentally alters the input token sequence and its positional encoding, shifting the "computational path" in high–dimensional representation space Naveed et al. (2024). Such shifts can increase sensitivity to the most recently presented quantifier/negation, activate different expert combinations in MoE architectures, or trigger different decision thresholds on logically ambiguous boundary cases. For example, a model might confuse the relationship between "No S are P" and "Some S are not P" during parsing, and a different premise order could alter this parsing path, thereby amplifying such errors. Therefore, inconsistency in LLM reasoning reflects an inherent limitation of their statistical learning paradigm when confronted with stringent symbolic operations.

Implications and Outlook

This study identifies a decoupling between accuracy and logical consistency, with logical consistency constituting an independent dimension. For example, Qwen3-235B ranks among the top in accuracy (92.19%) yet exhibits high logical inconsistency ($|\Delta\text{acc}| = 9.38\%$), whereas Doubao-Seed-1.6 shows moderate accuracy (78.12%) but good consistency. Moreover, merely using overall scores or a small number of examples is insufficient to identify weak spots like the EII and AAA–4 forms. Evaluation designs must therefore systematically cover the formal dimensions and provide stratified reporting. Benchmarks for LLMs should systematically incorporate tests of logical invariance, treating logical consistency as a core metric of equal importance to accuracy.

In addition, this study offers a refined conception and direction for "logical alignment." It is essential to clarify that the goal of logical alignment is adherence to objective logical axioms and theorems, rather than imitation of systematic human biases in reasoning. Our data show that the order sensitivity patterns of LLMs are almost entirely uncorrelated with human preference patterns (Pearson's $r = .052$, $p = .775$; Spearman's $\rho = 0.023$, $p = 0.899$). This near-zero correlation implies that simply making models "more human-like" by mimicking their average behavior would be a misguided approach to solving their logical fragility. Therefore, achieving genuine logical alignment requires the exploration of new technical pathways. A system that provides self-contradictory answers to logically equivalent inputs behaves in an unpredictable manner. Such inconsistency may be viewed as a form of "logical hallucination," rooted not in a lack of knowledge but in intrinsic deficiencies of the reasoning process. Addressing this fundamental issue is a necessary step toward reliable and genuinely general artificial intelligence.

Limitations and Future Directions

Although this study offers several perspectives and viewpoints, it also has certain limitations. The data for this study were collected using natural language in Chinese and within the content domain of plants/gardening; therefore, extrapolation to other languages and domains requires further validation. We used accuracy as the sole metric and did not collect response times, confidence ratings, or explanatory text. This choice preserves the purity of the logical evaluation but constrains finer–grained analysis of human and model processing. At the per–form level, no effects for the models reached the $p < .05$ significance threshold, which is why we used an effect size threshold ($|\Delta\text{textacc}| \geq 5\%$) to showcase the "top order effects." A larger item set or stricter sampling may detect weak yet consistent directional patterns.

Furthermore, this study was primarily conducted as a behavioral-level, black–box test. Future research could combine our approach with mechanistic interpretability techniques (e.g., circuit analysis) to "open the black box" and

investigate which specific neurons or attention heads contribute to the models' inconsistent behavior.

Finally, the LLMs in this study were all evaluated in a single–pass setting. Exploring whether prompting techniques such as multi-turn interaction, Chain-of-Thought (CoT), or self-correction can mitigate the models' reasoning fragility would be a valuable direction for future work.

Related Work

Syllogistic Reasoning Research in LLMs

Recent work has leveraged syllogisms as a classical probe for LLM reasoning.To analyze fine-grained biases, researchers have released dedicated benchmarks such as NeuBAROCO (Ozeki et al. 2024) and BIS Reasoning 1.0 (Japanese) (Nguyen et al. 2025), both of which increase coverage of figures, moods and linguistic variations.At the architectural level, Abbe et al. (2024) proposed the "Globality Barrier," while Saraipour and Zhang (2025) performed circuit analysis to locate the subnetworks that implement simple categorical inferences.Zong and Lin (2024) showed that many public datasets under-sample difficult figures and moods; Wang and Shi (2025) demonstrated that logical form predicts model behavior above and beyond n-gram likelihood.Extending beyond option ordering, Chen et al. (2024) found that re-ordering independent premises can dramatically affect accuracy on a variety of mathematical and logical tasks. He et al. (2025) mitigated this with Order-Centric Augmentation, randomly shuffling premises during training to improve robustness.

Research on Human Syllogistic Reasoning

Classical cognitive studies have documented multiple "formal effects"—systematic variations in difficulty induced by quantifier type, mood, and figure. For instance, the long-established Figure Effect was quantified by Dickstein (1978) and replicated in recent large-scale Chinese experiments (Jiang and Du 2022, 2024). These works also showed that figure biases the preferred term order of the conclusion (e.g., S–P vs. P–S). Additional formal factors include quantifier polarity and existential import (Jiang 2022).

Beyond form, content interacts with logic. The Belief Bias (Evans, Barston, and Pollard 1983) reveals that prior knowledge modulates acceptance of logically valid or invalid conclusions. Contemporary theories such as Mental Model Theory (Johnson-Laird 1983) and Dual-Process Theory (Evans 2003) aim to unify these observations by positing representational limits and heuristic–analytic trade-offs in human reasoning.

These human findings offer two benchmarks for LLM evaluation: overall accuracy relative to formal logic, and the pattern of errors predicted by figure, mood, and content. Our study draws on this tradition to build a fully balanced 64-item syllogistic set and to compare human and model performance under a tightly controlled premise-order manipulation.

Positioning and Distinctions of This Study

In the context of the syllogistic paradigm, which is logically insensitive to premise order, this study systematically examines the "order invariance" of LLMs in response to logically equivalent inputs, using accuracy as the sole metric.

Compared to previous work, our study's distinctions and contributions are threefold:

- **Logical Symmetry:** We only alter the presentation order of the two premises (forward/reverse) while keeping the logical form (figure $\times$ mood) identical. This isolates the "order effect" as a pure implementational difference under conditions of logical equivalence.
- **Direct Human-Machine Comparison:** We report the accuracy results and a model leaderboard for a large human sample (N=1317, on a random subset of 32 items) and 12 LLMs (on all 64 items) using the exact same dual-order test set. This approach avoids the inconsistencies caused by cross-dataset comparisons.
- **Fine-Grained Analysis:** At the level of the figures and moods, we identify systematically weak forms (e.g., negative–existential combinations; weak moods such as AAA-4) and quantify model-specific differences in order invariance (forward vs. reverse accuracy gaps) on these forms.

Our goal is not to explain the psychological mechanisms underlying the biases in humans or models. Instead, proceeding from the perspective of logical form and accuracy-based evaluation, we aim to precisely distinguish to what extent the behavioral patterns of models are "human-like" and to what extent they exhibit unique, non-human, machine-specific characteristics.

Conclusion

Within the classic paradigm of categorical syllogisms, this study systematically compared the reasoning consistency and robustness of 12 Large Language Models against a human–baseline ($N = 1317$) under premise order permutation, using accuracy as the primary metric. Our research confirms that while mainstream LLMs can achieve accuracy on formal reasoning tasks that far surpasses the average human, their reasoning processes generally fail to adhere to basic principles of logical invariance, exhibiting significant fragility and inconsistency.

We argue that humans and LLMs display two distinct patterns of "logical misalignment" in reasoning. Human reasoning errors manifest as predictable, mechanism–driven systematic biases; by contrast, LLM inconsistency appears as an unpredictable, computation–driven stochastic fluctuation, with behavior patterns largely unrelated to human cognitive preferences. This finding reveals the limitations of the current, predominantly accuracy–focused evaluation paradigm for LLMs and highlights the importance of "logical consistency" as an independent and core evaluation dimension. We argue that a system unable to ensure consistent answers to logically equivalent inputs may have misleadingly high accuracy and poses substantial risks in real–world applications, particularly in high–stakes domains.

References

Abbe, E.; Bengio, S.; Lotfi, A.; Sandon, C.; and Saremi, O. 2024. How Far Can Transformers Reason? The Globality Barrier and Inductive Scratchpad. arXiv:2406.06467.

Chen, X.; Chi, R. A.; Wang, X.; and Zhou, D. 2024. Premise Order Matters in Reasoning with Large Language Models. arXiv:2402.08939.

Dickstein, L. S. 1978. The effect of figure on syllogistic reasoning. *Memory & Cognition*, 6(1): 76–83.

Evans, J. S. B. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10): 454–459.

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.

He, Q.; He, Q.; Liang, J.; Xiao, Y.; Zhou, W.; Sun, Z.; and Yu, F. 2025. Order Doesn't Matter, But Reasoning Does: Training LLMs with Order-Centric Augmentation. arXiv:2502.19907.

Hou, Y.; Zhang, J.; Lin, Z.; Lu, H.; Xie, R.; McAuley, J.; and Zhao, W. X. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. arXiv:2305.08845.

Jiang, H. 2022. Effects of the Order and Type of Premises on Syllogistic Reasoning. *World Philosophy*, (02): 142–151.

Jiang, H.; and Du, G. 2022. Influence of Figure Effects of Syllogistic Reasoning on the Difficulty of Reasoning-An Analysis Based on an Investigation of 259 Junior Middle School Students. *Journal of Inner Mongolia Normal University(Educational Science Edition)*, 35(02): 83–89.

Jiang, H.; and Du, G. 2024. Comparison of the Figural Effect of Syllogistic Reasoning between Two Experimental Paradigms. *Psychological Exploration*, 44(01): 18–24.

Johnson-Laird, P. N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2024. A Comprehensive Overview of Large Language Models. arXiv:2307.06435.

Nguyen, H.-T.; Liu, C.; Liu, Q.; Tachibana, H.; Noe, S. M.; Miyao, Y.; Takeda, K.; and Kurohashi, S. 2025. BIS Reasoning 1.0: The First Large-Scale Japanese Benchmark for Belief-Inconsistent Syllogistic Reasoning. arXiv:2506.06955.

Ozeki, K.; Ando, R.; Morishita, T.; Abe, H.; Mineshima, K.; and Okada, M. 2024. Exploring Reasoning Biases in Large Language Models Through Syllogism: Insights from the NeuBAROCO Dataset. arXiv:2408.04403.

Pezeshkpour, P.; and Hruschka, E. 2023. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. arXiv:2308.11483.

Ranaldi, L.; and Zanzotto, F. 2024. HANS, are you clever? Clever Hans Effect Analysis of Neural Systems. In Bollegala, D.; and Shwartz, V., eds., *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, 314–325. Mexico City, Mexico: Association for Computational Linguistics.

Russell, S. J.; and Norvig, P. 2020. *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition.

Saraipour, K.; and Zhang, S. 2025. From Indirect Object Identification to Syllogisms: Exploring Binary Mechanisms in Transformer Circuits. arXiv:2508.16109.

Sumita, Y.; Takeuchi, K.; and Kashima, H. 2025. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 1009–1011.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926.

Wang, Y.; Cai, Y.; Chen, M.; Liang, Y.; and Hooi, B. 2024. Primacy Effect of ChatGPT. arXiv:2310.13206.

Wang, Y.; and Shi, F. 2025. Logical forms complement probability in understanding language model (and human) performance. arXiv:2502.09589.

Wu, M.; and Aji, A. F. 2023. Style Over Substance: Evaluation Biases for Large Language Models. arXiv:2307.03025.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. arXiv:2309.03882.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zong, S.; and Lin, J. 2024. Categorical Syllogisms Revisited: A Review of the Logical Reasoning Abilities of LLMs for Analyzing Categorical Syllogism. arXiv:2406.18762.

Appendix

Model Name	Developer/ Organization	Model Type/ Architecture	Scale/Level
GLM-4.5	Tsinghua University & Zhipu AI	Closed-source API (Transformer)	Flagship
Seed-OSS-36B-Instruct	DeepSeek	Open-source dense model (Transformer)	˜36B
Llama-4-Scout-17B-16E	Meta	Third-party MoE derivative based on the Llama family	16 experts $\times$ ˜1.06B
DeepSeek-V3	DeepSeek	Closed-source API (general-purpose / reasoning-enhanced)	Flagship
Qwen3-235B-A22B-Instruct	Alibaba	MoE with ˜235B total and ˜22B active parameters; coexists as both an API and an open-source model	235B (with 22B experts)
DeepSeek-R1	DeepSeek	Closed-source API (pure text optimization)	Flagship
Ling-max-1.5-0527	Shanghai AI Laboratory	Open-source dense model (Transformer)	˜30B
Claude-3.7-Sonnet	Anthropic	Closed-source API (Constitutional AI)	Flagship
Gemini-2.5-Pro	Google	Closed-source API (multimodal)	Flagship
GPT-4o	OpenAI	Closed-source API (end-to-end multimodal)	Flagship
Qwen3-8B	Alibaba	Open-source dense model (Transformer)	˜8B
Moonshot-v1-8k	Moonshot AI	Closed-source API (8k-context version)	˜8B

Table S1: Overview of the 12 Large Language Models Evaluated (Simplified).

Format	Accuracy (%)		Δ (%)	Sample Size		χ^2	*p*	Sig.
	Forward	Reverse		Fwd	Rev			
AAA-1	83.43	79.12	4.30	700	685	3.9409	0.0471	*
AAA-4	21.00	14.16	6.84	700	685	10.6921	0.0011	**
AAI-1	32.43	26.72	5.71	700	685	5.1520	0.0232	*
AAI-3	32.43	29.34	3.09	700	685	1.4025	0.2363	
AAI-4	49.00	47.15	1.85	700	685	0.4019	0.5261	
AAO-1	76.57	73.72	2.85	700	685	1.3563	0.2442	
AEE-2	36.14	42.04	-5.90	700	685	4.8197	0.0281	*
AEE-4	63.86	62.48	1.38	700	685	0.2255	0.6349	
AEI-2	50.07	50.00	0.07	685	700	0.0000	1.0000	
AEO-2	26.86	23.50	3.35	700	685	1.8915	0.1690	
AEO-4	30.66	34.43	-3.77	685	700	2.0743	0.1498	
AII-1	81.14	84.53	-3.38	700	685	2.5511	0.1102	
AII-3	49.00	48.03	0.97	700	685	0.0947	0.7583	
AIO-3	47.45	49.71	-2.27	685	700	0.6256	0.4290	
AOO-2	75.43	75.62	-0.19	700	685	0.0004	0.9836	
EAE-1	81.46	84.00	-2.54	685	700	1.3916	0.2381	
EAE-2	74.89	69.43	5.46	685	700	4.8697	0.0273	*
EAI-1	72.85	86.00	-13.15	685	700	35.9428	0.0000	***
EAI-3	67.29	67.01	0.28	700	685	0.0028	0.9577	
EAO-1	21.90	24.14	-2.25	685	700	0.8618	0.3532	
EAO-2	25.40	33.57	-8.17	685	700	10.7156	0.0011	**
EAO-3	35.18	37.57	-2.39	685	700	0.7535	0.3854	
EAO-4	32.41	37.71	-5.31	685	700	4.0488	0.0442	*
EII-2	30.86	26.13	4.73	700	685	3.5644	0.0590	
EII-4	69.78	52.71	17.07	685	700	41.7342	0.0000	***
EIO-1	76.20	72.57	3.63	685	700	2.2103	0.1371	
EIO-2	71.97	77.00	-5.03	685	700	4.3498	0.0370	*
EIO-3	43.94	42.71	1.23	685	700	0.1653	0.6843	
EIO-4	44.67	46.43	-1.76	685	700	0.3630	0.5468	
IAI-3	62.43	66.86	-4.43	700	685	2.7848	0.0952	
IAI-4	57.71	60.58	-2.87	700	685	1.0640	0.3023	
OAO-3	65.11	66.86	-1.75	685	700	0.3965	0.5289	

Table S2: Per-form syllogistic reasoning accuracy and statistical tests for human participants. Fwd = Forward order (Major premise first), Rev = Reverse order (Minor premise first). Δ = Forward accuracy - Reverse accuracy. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Sample sizes vary slightly due to excluded invalid responses

Statistic	Value	Unit
Overall Forward Accuracy Mean	52.80	%
Overall Forward Accuracy SD	20.27	%
Overall Reverse Accuracy Mean	52.87	%
Overall Reverse Accuracy SD	20.75	%
Mean Accuracy Difference (Δ)	-0.07	%
SD of Accuracy Differences	5.39	%
95% CI Lower Bound for Δ	-1.94	%
95% CI Upper Bound for Δ	1.79	%
Paired t-Statistic	-0.0771	
p-value (Paired t-test)	0.939	
Number of Syllogistic Formats	32	

Table S3: Overall Performance Statistics for Syllogistic Reasoning (Human Participants). SD = Standard Deviation, CI = Confidence Interval. Δ = Forward Accuracy - Reverse Accuracy. All accuracy values represent percentage correct responses. Statistical tests based on N = 32 syllogistic formats.

Category	Count	Percent
Both Forward Preference	16	50
Both Reverse Preference	16	50
No Difference	0	0
Sign Test p-value	1.000	
Total Formats	32	

Table S4: Direction preference analysis. Forward = Higher accuracy in forward order. Reverse = Higher accuracy in reverse order. Sign test shows no systematic bias (p = 1.000)

Format	Δ (%)	p-value	Sig
AAA-4	12.50	0.7029	
AEI-2	12.50	0.5930	
AEO-2	-12.50	0.6506	
AEO-4	12.50	0.7216	
EAO-3	12.50	0.5930	

Table S5: Notable order-effect formats ($p < 0.20$). Only formats with $p < 0.20$ shown. Δ = Forward accuracy - Reverse accuracy.

Statistic	Value			Interpretation
	Pooled OR	95% CI Lower	95% CI Upper	
Mantel-Haenszel Combined OR	0.9965	0.9568	1.0378	No overall bias

Table S6: Mantel-Haenszel combined odds ratio for syllogistic reasoning order effects (human participants). OR = Odds Ratio, CI = Confidence Interval. Combined OR calculated using Mantel-Haenszel method for 32 syllogistic formats. Interpretation based on CI inclusion of 1.0 (no effect).

Rank	Model	Coeff.	SE	95% CI	t	p	Sig.	N
1	GLM-4.5	12.500	7.4460	[-2.09, 27.09]	1.6787	0.1033	†	32
2	Seed-OSS-36B-Instruct	6.250	4.3476	[-2.27, 14.77]	1.4376	0.1606		32
3	Llama-4-Scout-17B-16E-Instruct	6.250	4.3476	[-2.27, 14.77]	1.4376	0.1606		32
4	GPT-4o	6.250	4.3476	[-2.27, 14.77]	1.4376	0.1606		32
5	DeepSeek-V3	3.125	3.1250	[-3.00, 9.25]	1.0000	0.3251		32
6	Gemini-2.5-Pro	3.125	3.1250	[-3.00, 9.25]	1.0000	0.3251		32
7	Qwen3-8B	3.125	5.4705	[-7.60, 13.85]	0.5712	0.5720		32
8	DeepSeek-R1	0.000	4.4901	[-8.80, 8.80]	0.0000	1.0000		32
9	Ling-max-1.5-0527	0.000	0.0000	[0.00, 0.00]	-	-		32
10	Moonshot-v1-8k	−3.125	8.3815	[-19.55, 13.30]	−0.3728	0.7118		32
11	Claude-3.7-Sonnet	−6.250	4.3476	[-14.77, 2.27]	−1.4376	0.1606		32
12	Qwen3-235B-A22B-Instruct-2507	−9.375	5.2351	[-19.64, 0.89]	−1.7908	0.0831	†	32

Table S7: Model-level order-effect coefficients. Coeff. = Order-effect coefficient (%); SE = Standard Error; CI = Confidence Interval; N = Number of Formats. Significance levels: † p ¡ 0.10, * p ¡ 0.05.

Rank	Model	Pooled OR	95% CI	z	p	Sig.	Interpretation
1	GLM-4.5	1.3478	[0.6313, 2.8774]	0.7714	0.4405		Weak preference (n.s.)
2	Seed-OSS-36B-Instruct	1.1739	[0.5354, 2.5738]	0.4003	0.6889		Weak preference (n.s.)
3	Llama-4-Scout-17B-16E-Instruct	1.1739	[0.5354, 2.5738]	0.4003	0.6889		Weak preference (n.s.)
4	GPT-4o	1.1739	[0.5354, 2.5738]	0.4003	0.6889		Weak preference (n.s.)
5	DeepSeek-V3	1.0851	[0.4914, 2.3962]	0.2021	0.8399		No bias
6	Gemini-2.5-Pro	1.0851	[0.4914, 2.3962]	0.2021	0.8399		No bias
7	Qwen3-8B	1.0816	[0.4975, 2.3514]	0.1981	0.8430		No bias
8	DeepSeek-R1	1.0000	[0.4566, 2.1902]	0.0000	1.0000		No bias
9	Ling-max-1.5-0527	1.0000	[0.4493, 2.2259]	0.0000	1.0000		No bias
10	Moonshot-v1-8k	0.9298	[0.4402, 1.9640]	−0.1907	0.8488		No bias
11	Claude-3.7-Sonnet	0.8519	[0.3885, 1.8677]	−0.4003	0.6889		Weak preference (n.s.)
12	Qwen3-235B-A22B-Instruct-2507	0.7895	[0.3624, 1.7198]	−0.5951	0.5518		Weak preference (n.s.)

Table S8: Model-level pooled odds ratios. OR = Odds Ratio; CI = Confidence Interval; n.s. = not significant. OR ¿ 1 indicates a preference for the forward presentation order.

Format	Human			LLM			Difference	
	Fwd (%)	Rev (%)	Δ (%)	Fwd (%)	Rev (%)	Δ (%)	$\Delta_L - \Delta_H$	$\lvert\Delta_L - \Delta_H\rvert$
EII-4	69.78	52.71	17.07	31.25	31.25	0.00	−17.07	17.07
AEO-4	30.66	34.43	−3.77	62.50	50.00	12.50	16.27	16.27
AEO-2	26.86	23.50	3.35	75.00	87.50	−12.50	−15.85	15.85
EAO-3	35.18	37.57	−2.39	93.75	81.25	12.50	14.89	14.89
EAI-1	72.85	86.00	−13.15	100.00	100.00	0.00	13.15	13.15
AEI-2	50.07	50.00	0.07	93.75	81.25	12.50	12.43	12.43
AEE-2	36.14	42.04	−5.90	93.75	87.50	6.25	12.15	12.15
EII-2	30.86	26.13	4.73	12.50	18.75	−6.25	−10.98	10.98
AAI-3	32.43	29.34	3.09	87.50	93.75	−6.25	−9.34	9.34
IAI-4	57.71	60.58	−2.87	93.75	87.50	6.25	9.12	9.12
EAO-1	21.90	24.14	−2.25	87.50	81.25	6.25	8.50	8.50
EAO-2	25.40	33.57	−8.17	93.75	93.75	0.00	8.17	8.17
OAO-3	65.11	66.86	−1.75	100.00	93.75	6.25	8.00	8.00
EIO-3	43.94	42.71	1.23	93.75	100.00	−6.25	−7.48	7.48
EAI-3	67.29	67.01	0.28	68.75	62.50	6.25	5.97	5.97
AAA-4	21.00	14.16	6.84	37.50	25.00	12.50	5.66	5.66
EAO-4	32.41	37.71	−5.31	50.00	50.00	0.00	5.31	5.31
EIO-2	71.97	77.00	−5.03	100.00	100.00	0.00	5.03	5.03
IAI-3	62.43	66.86	−4.43	100.00	100.00	0.00	4.43	4.43
AAA-1	83.43	79.12	4.30	100.00	100.00	0.00	−4.30	4.30
EIO-1	76.20	72.57	3.63	100.00	100.00	0.00	−3.63	3.63
AII-1	81.14	84.53	−3.38	93.75	100.00	−6.25	−2.87	2.87
AAO-1	76.57	73.72	2.85	100.00	100.00	0.00	−2.85	2.85
EAE-1	81.46	84.00	−2.54	100.00	100.00	0.00	2.54	2.54
AIO-3	47.45	49.71	−2.27	56.25	56.25	0.00	2.27	2.27
AAI-4	49.00	47.15	1.85	50.00	50.00	0.00	−1.85	1.85
EIO-4	44.67	46.43	−1.76	93.75	93.75	0.00	1.76	1.76
AEE-4	63.86	62.48	1.38	100.00	100.00	0.00	−1.38	1.38
AII-3	49.00	48.03	0.97	100.00	100.00	0.00	−0.97	0.97
EAE-2	74.89	69.43	5.46	100.00	93.75	6.25	0.79	0.79
AAI-1	32.43	26.72	5.71	87.50	81.25	6.25	0.54	0.54
AOO-2	75.43	75.62	−0.19	93.75	93.75	0.00	0.19	0.19

Table S9: Comparison of order effects between humans and LLMs. All accuracy values are in percent (%). Fwd: Forward order. Rev: Reverse order. $\Delta = \text{Accuracy}_{\text{Fwd}} - \text{Accuracy}_{\text{Rev}}$. Δ_L is the delta for LLMs, Δ_H for Humans.