Consistent Biases in Large Language Models' Syllogistic Reasoning

Limeng Ge

East China Normal University
Department of Philosophy
lg3635@nyu.edu

Abstract

Large language models (LLMs) have demonstrated impressive performance across a wide range of language understanding tasks. However, whether they can truly reason logically over natural language remains an open question. While prior studies have evaluated LLMs on various formal logic benchmarks, a systematic mapping of their performance on categorical syllogisms remains under-explored, particularly regarding the interplay between logical structure and linguistic heuristics. To address this gap, we present a systematic evaluation of LLMs on categorical syllogistic reasoning, covering all four Figures and both valid and invalid Moods. We assess five representative models: GPT-4o, Gemini-2.0-Flash, LLaMA-3.3-70B, Qwen-3-Max, and DeepSeek-Chat. Our results reveal that LLMs exhibit limited formal reasoning ability and perform particularly poorly on invalid syllogisms, where linguistic plausibility conflicts with logical validity. Moreover, all models show a consistent bias toward the syntactic position of the middle term, suggesting that their reasoning relies on surface linguistic cues rather than abstract logical structures. We hope this work provides a foundation for more rigorous evaluation and improvement of logical reasoning in future language models.

Code — https://github.com/limengge426/llm-syllogism

Introduction

Syllogistic inferences, formalized since Aristotle's *Prior Analytics* (Jenkinson 1985) in the 4th century BC, represent one of the oldest and most extensively studied forms of deductive reasoning. While syllogisms constitute only a limited subset of predicate logic compared to comprehensive formal systems, they have experienced renewed attention in recent decades through natural logic research (van Benthem 1983; Sánchez 1991; Lappin and Fox 2015). This revival stems from recognizing syllogistic reasoning as a "natural" inference system applicable to everyday reasoning in ordinary language. Beyond their linguistic relevance, syllogisms serve as a benchmark across diverse disciplines, from cognitive psychology (Stenning and van Lambalgen 2008) and diagrammatic reasoning (Ando et al. 2023) to neuroscience (Goel et al. 1998), all drawing upon syllogistic forms as reference points for studying human inference.

The emergence of deep learning-based natural language AI tools, particularly state-of-the-art Large Language Models (LLMs) such as BERT (Devlin et al. 2019) and GPT (Brown et al. 2020), presents both remarkable opportunities and open questions for logical reasoning research. While these models demonstrate impressive capabilities in language understanding and generation tasks, their ability to perform rigorous logical inference remains an active area of investigation (Cheng et al. 2025; Liu et al. 2025). A fundamental question persists: Do LLMs genuinely reason, or do they primarily exploit statistical patterns to simulate reasoning capabilities? This question carries profound implications for deploying these systems in applications requiring robust logical inference.

However, there exists a significant gap between advances in LLM technology and insights from cognitive science research on human reasoning. The field of cognitive science has accumulated extensive knowledge about systematic patterns in human logical reasoning, including various biases and heuristics that characterize how humans approach deductive inference tasks (Evans 1990; Byrne, Evans, and Newstead 1993; Khemlani and Johnson-Laird 2012). Despite this rich foundation, these insights have not been fully integrated with contemporary AI research. Recent work has begun exploring whether LLMs exhibit reasoning patterns similar to those observed in humans (Eisape et al. 2024; Ando et al. 2023; Bertolazzi, Gatt, and Bernardi 2024), but critical questions remain unanswered.

First, existing studies have not systematically examined whether reasoning patterns represent architecture-independent regularities or are idiosyncratic to specific model families (Mondorf and Plank 2024). Do large language models trained on diverse architectures, datasets, and objectives converge on similar reasoning patterns, or do their biases differ in unpredictable ways? Second, while previous work has evaluated models on subsets of syllogistic forms (Ando et al. 2023), no study has comprehensively mapped LLM behavior across the complete space of valid and invalid syllogisms as systematically characterized in formal logic. Third, the mechanisms underlying observed reasoning patterns remain poorly understood (Jiang et al. 2024). Are models relying on genuine logical deduction, structural features of premises, or other heuristics?

In this paper, we address these gaps through a system-

atic evaluation of five representative LLMs—Qwen3-Max, DeepSeek-Chat, GPT-4o, Gemini-2.0-Flash, and LLaMA-3.3-70B—across all 44 valid and invalid syllogistic forms spanning 11 moods and 4 figures. Each syllogism was instantiated using semantically coherent relations drawn from WordNet (Miller 1992) to ensure content neutrality while maintaining linguistic naturalness. Critically, rather than comparing absolute accuracy levels, we analyze the consistency of reasoning patterns across models, examining whether diverse architectures exhibit similar difficulty gradients and error profiles.

Our experimental design enables investigation of three key questions: (1) Do reasoning difficulties remain consistent across architecturally diverse models? (2) Do models exhibit systematic asymmetries between valid and invalid syllogisms? (3) Does the structural arrangement of syllogistic premises systematically influence reasoning accuracy? By examining cross-model correlations in difficulty patterns and analyzing performance as a function of logical structure, we can distinguish between genuine logical reasoning and reliance on structural heuristics.

The contributions of this paper are as follows:

- We systematically evaluate five large language models on the full set of forty-four classical syllogistic forms.
- Our experiments reveal strong cross-model consistency in reasoning difficulty, indicating shared structural biases across architectures.
- We identify clear asymmetries between valid and invalid reasoning, and figure-dependent variations linked to the position of the middle term.

Taken together, these findings provide a structured account of how statistical language models approximate logical inference. The consistency of reasoning errors across diverse architectures implies that deviations from normative logic stem not from implementation details but from fundamental properties of language-based learning. The parallels with human reasoning further suggest that both artificial and human cognition may rely on heuristic alignment rather than strict formal deduction. These insights advance theoretical understanding of reasoning as an emergent property of linguistic modeling and highlight the importance of structural analysis for the safe and interpretable deployment of LLMs.

Background and Related Work

Syllogism

A syllogism is a form of deductive reasoning where a conclusion is inferred from two given premises, with all three statements being categorical propositions. A categorical proposition here means a simple quantified subject–predicate statement relating two terms. It typically takes one of four standard forms: A ("All S are P"), E ("No S are P"), I ("Some S are P"), and O ("Some S are not P"). In this paper, each premise and the conclusion is instantiated using these A/E/I/O templates. Every syllogism is composed of three terms: the major term (P), which is the predicate of the conclusion; the minor term (S), which is the subject of the conclusion; and the middle term (M), which appears in both premises but not in the conclusion. The logical structure of a syllogism is determined by its Figure and its Mood.

The Figure. The Figure of a syllogism is determined by the positions of the middle term (M) in the two premises. There are four possible figures of a syllogism (Table 1).

Figure 1	Figure 2	Figure 3	Figure 4
M – P	P – M	M – P	P – M
S – M	S – M	M – S	M – S
S – P	S – P	S – P	S – P

Table 1: Four types of figures.

The Mood. The Mood of a syllogism is defined by the type of categorical propositions (A, E, I, O) used for its major premise, minor premise, and conclusion. The four proposition types are shown in Table 2.

Type	Form	Description
A	All S are P	Universal affirmative
E	No S are P	Universal negative
I	Some S are P	Particular affirmative
O	Some S are not P	Particular negative

Table 2: Four types of moods.

For example, consider the following argument:

- Major Premise (A): All felines (M) are mammals (P).
- Minor Premise (A): All tigers (S) are felines (M).
- Conclusion (A): Therefore, all tigers (S) are mammals (P).

In terms of its Figure, this example belongs to the First Figure. In terms of its Mood, all three propositions are universal affirmative (A-type). Therefore, the logical form of this syllogism is AAA-1.

The Figure and Mood together define the logical form. Since there are 4 Figures and 16 possible Moods, there are 64 possible syllogistic forms. According to the rules of deduction, only 11 of these 16 Moods (AAA, AAI, AEE, AEO, AII, AOO, EAE, EAO, EIO, IAI, OAO) can yield a valid conclusion in at least one Figure. When distributed across the 4 Figures and tested with their specific rules, these Moods result in a total of 24 valid forms (Table 3).

Note that in our experiments, all S, M, and P terms are drawn from WordNet (Miller 1992). This ensures that the terms (e.g., dog, mammal) are non-empty and possess existential import. Therefore, the "conditionally valid forms" (i.e., those that require the existence of the subject, such as AAI-1 or EAO-3), which are sometimes debated in modern logic, are treated as unambiguously valid in our study.

Related Work

As large language models (LLMs) continue to evolve, it becomes increasingly crucial to evaluate their diverse reasoning capabilities (Cheng et al. 2025). A significant body

	AAA	AAI	AEE	AEO	AII	AOO	EAE	EAO	EIO	IAI	OAO
Figure 1	✓	(✓)			✓		✓	(✓)	✓		
Figure 2			✓	(✓)		✓	✓	(✓)	✓		
Figure 3		(✓)			✓			(✓)	✓	✓	✓
Figure 4		(✓)	✓	(✓)				(✓)	✓	✓	

Table 3: Twenty-four types of valid forms. Forms marked with parentheses are conditionally valid mood, as they assume existential import for the subject and predicate terms, making the inference valid only under this condition.

of work has emerged to benchmark logical reasoning, with pioneering datasets like LogicBench (Parmar et al. 2024) systematically evaluating a wide range of inference rules across propositional, first-order, and non-monotonic logics. Other works have focused on enhancing these capabilities by training on principled, synthetic logic corpora (Morishita et al. 2024). While this research provides a vital, broad-strokes understanding of LLM reasoning on formal logic, it has often focused on propositional and first-order rules, leaving classical categorical syllogistic reasoning, a natural-language-aligned and tightly controlled testbed for diagnosing deductive validity and systematic biases, comparatively less explored.

A parallel line of inquiry has begun to investigate this specific syllogistic reasoning ability, often with a focus on human cognitive parallels. Foundational studies in this area, such as those by Ando (Ando et al. 2023), used the NeuBAROCO dataset to demonstrate that models like GPT-3.5 struggle with syllogisms and exhibit human-like biases, including belief bias, conversion errors, and atmosphere effects. Similarly, Eisape et al. (Eisape et al. 2024) provided a systematic comparison of the PaLM 2 model family with human reasoners, finding that even the largest models make systematic errors that mirror human fallacies. Bertolazzi et al. (Bertolazzi, Gatt, and Bernardi 2024) further solidified this by explicitly connecting the behavior of LLaMA-3 to the Atmosphere Heuristic Theory (Woodworth and Sells 1935). This prior work has been essential in establishing that LLMs, like humans, rely on cognitive shortcuts for syllogistic tasks.

While these studies have established the existence of such biases, several crucial attributes motivated our present study. First, existing work has often focused on one or two model families (e.g., PaLM 2 or LLaMA-3) or has not systematically compared a diverse set of state-of-the-art, architecturally distinct models on an identical task. This leaves it unclear whether these biases are idiosyncratic artifacts or a robust, architecture-independent property of all LLMs. Second, no single study has provided a comprehensive map of LLM performance across the complete space of all 44 valid and invalid syllogistic forms, which is necessary to fully diagnose the structure of their failure. Our work addresses this gap by providing a systematic, cross-model evaluation of five diverse LLMs. Rather than just identifying the presence of biases, we measure the consistency of these bias patterns across all models, allowing us to distinguish model-specific quirks from the fundamental, "pseudo-logical" heuristics that all current LLMs may convergently learn.

Method and Experiment

Task Definition

We move beyond simple classification of entire syllogisms by defining our core task as Conclusion Correctness Verification. In this task, the model receives two premises as a contextual input and a separate conclusion, and must determine whether the conclusion logically follows from the premises.

This design enables a rigorous and tightly controlled evaluation of deductive reasoning by testing whether a model can both accept valid inferences and reject invalid ones. We evaluate models on two primary sample types:

Type 1: Validity Test. For each of the 24 logically valid forms, we pair the premises with their correctly entailed conclusion. A logically competent model is expected to answer "Yes", confirming the valid inference. For example, testing the valid form AAA-1:

- Context: All engines are machines. All turbines are engines.
- Conclusion: All turbines are machines.
- Expected: Yes

Type 2: Invalidity Test. For this test, we pair premises with a conclusion that is not logically entailed. A robust model is expected to answer "No", identifying the non-sequitur. For each of the 20 logically invalid forms, we present their corresponding fallacious conclusion. For example, testing the invalid form OAO-4:

- Context: Some furniture are not seats, All seats are chairs.
- Conclusion: Some chairs are not furniture.
- Expected: No

This two-pronged evaluation (Validity and Invalidity) is critical. It stresses that true logical competence requires not only the ability to affirm correct deductions (Type 1) but also the equally important and more challenging ability to reject non-entailed conclusions (Type 2).

Benchmark Construction

We utilized WordNet (Miller 1992), a large lexical database of English, to generate semantically coherent and factually plausible test items. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), which are then linked by conceptual-semantic relations. Critically for our work, this includes

the hyponym/hypernym ('is-a') hierarchy, which provides a structured source of semantic relationships.

The construction of natural language instances for this task began with establishing a logical foundation covering all 44 classical syllogistic forms (24 valid and 20 invalid). To ensure semantic coherence while isolating logical form from world knowledge, we instantiated these forms using term triplets drawn from WordNet's hyponym–hypernym hierarchy.

A primary challenge in this evaluation is the "Belief Bias," where a model might incorrectly accept a fallacious conclusion simply because its content is factually believable. To isolate the assessment of formal structure from world knowledge, we populated these 44 logical templates using semantic relations drawn exclusively from the WordNet lexical database.

We traversed the WordNet hyponym/hypernym hierarchy to extract semantically coherent (S, M, P) term triplets. For example, a chain like 'spaniel' (S) $\rightarrow$ 'dog' (M) $\rightarrow$ 'mammal' (P) provides a factually true and linguistically natural basis for instantiating premises such as 'All spaniels are dogs' and 'All dogs are mammals.' This methodology was used to generate 250 unique instances for each logical form. For the Validity Test (Type 1), we generated 250 instances for each of the 24 valid forms, yielding 6,000 test items where the expected answer is "Yes". For the Invalidity Test, we generated 250 instances for each of the 20 logically invalid forms. This resulted in a total of 5,000 invalid test items. This process yielded a final benchmark of 11,000 test items, where the logical form, not the content, was the critical variable under evaluation.

Experiment and Evaluation

We evaluated five representative LLMs on the benchmark: GPT-4o (OpenAI 2024), Gemini-2.0-Flash (Google DeepMind 2024), LLaMA-3.3-70B (Grattafiori et al. 2024), Qwen3-Max (Yang et al. 2024), and DeepSeek-Chat (DeepSeek-AI et al. 2024).

To execute the Conclusion Correctness Verification task, we used a standardized, direct-answer prompting strategy. Each model was presented with the two premises as context and asked to evaluate the conclusion in a zero-shot manner. We deliberately avoided Chain-of-Thought (CoT) prompting to probe the models' default, intuitive reasoning heuristics rather than their ability to follow an explicit reasoning script.

Our evaluation is twofold. First, we measure Accuracy as the primary performance metric, calculated separately for each of the two test types to assess model correctness. Second, to test our hypothesis of architecture-independent bias, we measure Cross-Model Consistency. We compute the Pearson correlation coefficient (r) between the accuracy vectors (one vector of 44 accuracy scores per model, one for each syllogistic form) for every pair of models. A high positive correlation indicates that different models find the same logical forms systematically easy or difficult, suggesting a convergent, underlying bias.

Results and Analysis

Overall results

Table 4 shows the accuracy scores for all models across the 44 syllogistic forms.

DeepSeek-Chat The overall accuracy was high (84.03%). The model's performance was relatively stable across the different logical structures (Figures). Performance was highest on Figure 3 (86.98%), while performance was lowest on Figure 1 (80.78%).

Gemini-2.0-Flash The overall accuracy for Gemini-2.0-Flash was 69.91%. The model answered most correctly on Figure 3 (77.52%), but less on Figure 1 (64.49%) and still less on Figure 2 (62.82%).

GPT-4o The overall accuracy was 75.40%. A clear performance difference was observed across Figures. The model performed best on Figure 3 (81.47%), although accuracy was significantly lower on Figure 1 (67.34%).

LLaMA-3.3-70B The overall accuracy was 71.79%. The results show a strong dependency on the Figure. The model's accuracy was highest on Figure 3 (81.54%), but it struggled significantly with Figure 2, where accuracy dropped to its lowest point (58.70%).

Qwen3-Max The results on Qwen3-Max show a similarly high performance. The overall accuracy was 81.83%. The model also demonstrated sensitivity to the syllogism's Figure, performing best on Figure 3 (87.58%) and worst on Figure 2 (74.62%).

Cross-Model Consistency

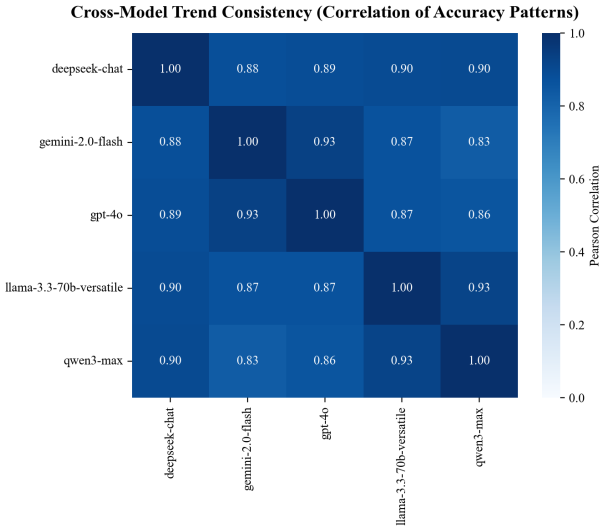

Fig. 1: Cross-Model Trend Consistency

We analyze cross-model consistency in syllogistic reasoning. Each model's performance matrix—comprising accuracies across 44 syllogistic forms—was flattened into a 44-dimensional vector that characterizes its reasoning profile. We then computed pairwise Pearson correlation coefficients between these vectors, producing a 5×5 correlation matrix

	Figure/Mood	AAA	AAI	AEE	AEO	AII	AOO	EAE	EAO	EIO	IAI	OAO	Average
Deepseek-chat	Figure 1	1.000	1.000	**0.856**	**0.649**	1.000	**0.852**	0.986	0.911	0.956	**0.152**	**0.525**	0.808
	Figure 2	**0.720**	**0.661**	0.947	0.939	**0.650**	0.957	0.934	0.926	0.922	**0.498**	**0.832**	0.817
	Figure 3	**0.381**	1.000	**0.916**	**0.784**	1.000	**1.000**	**0.530**	0.998	0.983	1.000	0.976	**0.870**
	Figure 4	**0.816**	0.996	0.983	0.931	**0.437**	**0.904**	**0.672**	0.955	0.948	0.996	**0.892**	0.866
Gemini-2.0-flash	Figure 1	1.000	1.000	**0.343**	**0.136**	1.000	**0.422**	0.956	0.939	0.991	**0.026**	**0.283**	0.645
	Figure 2	**0.473**	**0.167**	0.957	0.941	**0.174**	0.932	0.824	0.865	0.836	**0.078**	**0.664**	0.628
	Figure 3	**0.325**	1.000	**0.809**	**0.280**	1.000	**0.953**	**0.231**	0.984	0.962	1.000	0.983	**0.775**
	Figure 4	**0.905**	0.996	0.993	0.942	**0.092**	**0.647**	**0.443**	0.927	0.878	0.996	**0.411**	0.748
GPT-4o	Figure 1	1.000	1.000	**0.312**	**0.172**	1.000	**0.440**	0.969	0.961	0.974	**0.069**	**0.512**	0.673
	Figure 2	**0.663**	**0.504**	0.906	0.800	**0.255**	0.924	0.941	0.923	0.872	**0.455**	**0.830**	0.734
	Figure 3	**0.324**	1.000	**0.797**	**0.627**	1.000	**1.000**	**0.363**	0.980	0.921	1.000	0.950	**0.815**
	Figure 4	**0.802**	0.996	0.993	0.873	**0.169**	**0.931**	**0.424**	0.883	0.877	0.996	**0.791**	0.794
LLaMA-3.3-70b	Figure 1	1.000	1.000	**0.674**	**0.290**	1.000	**0.651**	0.804	0.815	0.928	**0.026**	**0.460**	0.695
	Figure 2	**0.510**	**0.232**	0.758	0.829	**0.291**	0.757	0.656	0.664	0.714	**0.106**	**0.940**	0.587
	Figure 3	**0.208**	1.000	**0.925**	**0.577**	1.000	**0.972**	**0.557**	0.930	0.827	1.000	0.974	**0.815**
	Figure 4	**0.649**	0.996	0.944	0.854	**0.069**	**0.918**	**0.755**	0.769	0.738	0.996	**0.825**	0.774
Qwen3-max	Figure 1	1.000	1.000	**0.660**	**0.640**	1.000	**0.796**	0.986	0.948	0.970	**0.101**	**0.768**	0.806
	Figure 2	**0.594**	**0.493**	0.959	0.955	**0.496**	0.932	0.868	0.789	0.800	**0.382**	**0.940**	0.746
	Figure 3	**0.385**	0.998	**0.953**	**0.873**	1.000	**1.000**	**0.566**	0.984	0.899	1.000	0.976	**0.876**
	Figure 4	**0.896**	0.972	0.993	0.700	**0.405**	**0.960**	**0.840**	0.906	0.716	0.980	**0.929**	0.845
Average		0.683	0.850	0.834	0.690	0.652	0.847	0.715	0.903	0.886	0.593	0.773	

Table 4: Accuracy (%) of five LLMs across 44 classical syllogistic forms and four figures. Bolded cells indicate invalid syllogisms, while bolded averages mark the best-performing figure for each model.

representing the similarity of reasoning trends among models. Pearson's correlation coefficient (Evans, Barston, and Pollard 1983) quantifies the linear association between two distributions, capturing whether models exhibit similar relative difficulty patterns even when their absolute accuracies differ.

As illustrated in Fig. 1, the average cross-model correlation reaches $r_{\text{mean}} = 0.886$, indicating a high level of alignment across all five models. Every pairwise correlation exceeds 0.83, suggesting that despite differences in architecture, scale, and training corpora, the models converge toward highly similar reasoning structures. This convergence implies that LLMs may internalize comparable inductive heuristics when processing syllogistic relations, rather than reasoning purely from symbolic logic. Such regularities may emerge from shared statistical properties of natural language, where co-occurrence patterns between quantifiers and logical connectives shape the models' internal representations of logical structure. In this sense, the consistency observed across architectures echoes findings from cognitive psychology, suggesting that both humans and LLMs may rely on analogous structural heuristics when performing deductive reasoning.

Middle-Term Bias

As discussed in Section Syllogism, the four classical Figures in syllogistic logic differ solely in the position of the middle term (M) across the two premises. This structural variation determines whether M functions as the grammatical subject or predicate, and thereby governs the inferential pathway linking the subject (S) and predicate (P) of the conclusion. While this distinction has long been noted in human reasoning studies, we find that large language models exhibit a similar sensitivity to the position of the middle term.

As shown in Table 5, model performance follows a clear gradient across the four Figures: Figure 3 achieves the highest average accuracy (0.830), followed by Figure 4 (0.806), whereas Figure 1 (0.726) and Figure 2 (0.702) are considerably lower. This pattern aligns precisely with the number of times the middle term serves as the subject across the two premises: in Figure 3 (M-P, M-S), M is the subject twice; in Figure 4 (P-M, M-S) and Figure 1 (M-P, S-M), once; and in Figure 2 (P-M, S-M), never. The monotonic progression suggests a robust middle-term bias: the more frequently M occupies the subject position, the more likely the model is to draw a correct conclusion.

Figure	**Accuracy (%)**
Figure 1	72.65
Figure 2	70.24
Figure 3	83.01
Figure 4	80.55

Table 5: Accuracy (%) of four types of figures.

Model	Accuracy on Valid Syllogisms (%)	Accuracy on Invalid Syllogisms (%)
DeepSeek-Chat	92.5	58.6
Gemini-2.0-Flash	83.4	48.7
GPT-4o	88.6	62.3
LLaMA-3.3-70B	85.3	56.2
Qwen3-Max	89.2	63.1

Table 6: Accuracy (%) of the models on valid and invalid syllogisms.

Valid-Invalid Asymmetry

Across all five models, a clear asymmetry emerges between valid and invalid syllogisms. As shown in Table 6, the average accuracy on valid syllogisms (0.93) substantially exceeds that on invalid ones (0.56). This disparity indicates that LLMs tend to accept conclusions even when they are not logically entailed by the premises, which is a phenomenon analogous to the belief-bias and acceptance-bias effects documented in human syllogistic reasoning. Rather than explicitly verifying logical entailment, models appear to rely on surface-level associations between premise and conclusion terms, leading to high confidence in syntactically plausible but logically invalid inferences.

Analysis and Discussion

The findings above show that large language models, although trained with different architectures and datasets, converge toward strikingly similar reasoning profiles when performing categorical syllogistic reasoning. This convergence suggests that the observed patterns—cross-model consistency, middle-term bias, and valid–invalid asymmetry—reflect deep regularities in how statistical language models approximate reasoning. Unlike formal logical systems that manipulate explicit symbols and rules, language models rely on the probabilistic associations that emerge from linguistic co-occurrence patterns. Quantifiers such as all, some, as well as negation tend to appear in stable syntactic configurations across texts, and such regularities may induce latent templates that imitate logical reasoning without guaranteeing logical validity.

Within this perspective, the middle-term bias can be interpreted as a by-product of linguistic structure. When the middle term (M) serves as the subject of both premises, the syntactic direction of inference ("M $\rightarrow$ P" and "M $\rightarrow$ S") aligns with the dominant forward entailment pattern in natural language. When M functions as the predicate, however, the inferential relation must reverse ("S $\rightarrow$ M" and "M $\rightarrow$ P"), a configuration that is rarer and less intuitive for models trained primarily on sequential dependencies. This pattern implies that models favor subject-position M not because of logical strategy but because of the statistical distribution of entailment in human language.

The valid–invalid asymmetry further illustrates this linguistic dependence. Across all models, semantically coherent premises such as "Some animals are mammals; All dogs are animals; Therefore, some dogs are mammals" often lead to acceptance of invalid conclusions. This tendency reveals that models assess plausibility rather than logical entailment, echoing the "atmosphere effect" originally described by Woodworth (Woodworth and Sells 1935) and the belief-bias phenomenon in human syllogistic reasoning (**?**). In both humans and models, judgment is guided by heuristic cues derived from the semantic or quantitative "atmosphere" of the premises rather than by strict deductive form.

Together these results point to a structural tension between linguistic generalization and logical generalization. Large language models reproduce the form of reasoning while failing to internalize its function. They capture the surface regularities that correlate with validity in ordinary language but collapse when those correlations are broken. Recognizing this boundary between linguistic inference and formal logic clarifies both the cognitive parallels between humans and large language models and the fundamental challenges in building systems capable of genuine deductive reasoning.

Conclusion

In this work, we evaluated large language models on formal syllogistic reasoning across 44 canonical logical forms. Using a WordNet-based dataset and a unified prompt framework, we systematically assessed five representative models including GPT-4o, Gemini-2.0-Flash, DeepSeek-Chat, Qwen3-Max, and LLaMA-3.3-70B. The results reveal consistent reasoning patterns across architectures, with strong cross-model agreement and clear figure-dependent differences. Models perform better when the middle term appears as the grammatical subject and show a marked tendency to accept invalid but semantically coherent conclusions. These findings suggest that large language models rely on linguistic heuristics rather than strict logical deduction, capturing the surface form of reasoning without fully internalizing its formal structure. Future progress may depend on integrating explicit symbolic mechanisms with language-based reasoning systems.

There remain many issues to be addressed. First, while prompting strategies like Chain-of-Thought (CoT) are known to boost accuracy, it remains an open question to what extent they can correct the middle-term biases we identify, or if they merely mask these flawed heuristics. Second, given that LLMs and humans exhibit analogous failures, parallel experiments would be valuable to explore whether this stems from shared representational mechanisms or mere functional convergence. Finally, it is interesting to consider if the syntactic biases we found extend beyond classical syllogisms into more complex domains of natural logic, such

as reasoning with generalized quantifiers.

References

Ando, R.; Morishita, T.; Abe, H.; Mineshima, K.; and Okada, M. 2023. Evaluating Large Language Models with NeuBAROCO: Syllogistic Reasoning Ability and Human-like Biases. arXiv:2306.12567.

Bertolazzi, L.; Gatt, A.; and Bernardi, R. 2024. A Systematic Analysis of Large Language Models as Soft Reasoners: The Case of Syllogistic Inferences. arXiv:2406.11341.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Byrne, R. M. J.; Evans, J. S. B. T.; and Newstead, S. E. 1993. *Human Reasoning: The Psychology of Deduction*. Psychology Press, 1st edition.

Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering LLMs with Logical Reasoning: A Comprehensive Survey. arXiv:2502.15652.

DeepSeek-AI; :; Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; Gao, H.; Gao, K.; Gao, W.; Ge, R.; Guan, K.; Guo, D.; Guo, J.; Hao, G.; Hao, Z.; He, Y.; Hu, W.; Huang, P.; Li, E.; Li, G.; Li, J.; Li, Y.; Li, Y. K.; Liang, W.; Lin, F.; Liu, A. X.; Liu, B.; Liu, W.; Liu, X.; Liu, X.; Liu, Y.; Lu, H.; Lu, S.; Luo, F.; Ma, S.; Nie, X.; Pei, T.; Piao, Y.; Qiu, J.; Qu, H.; Ren, T.; Ren, Z.; Ruan, C.; Sha, Z.; Shao, Z.; Song, J.; Su, X.; Sun, J.; Sun, Y.; Tang, M.; Wang, B.; Wang, P.; Wang, S.; Wang, Y.; Wang, Y.; Wu, T.; Wu, Y.; Xie, X.; Xie, Z.; Xie, Z.; Xiong, Y.; Xu, H.; Xu, R. X.; Xu, Y.; Yang, D.; You, Y.; Yu, S.; Yu, X.; Zhang, B.; Zhang, H.; Zhang, L.; Zhang, L.; Zhang, M.; Zhang, M.; Zhang, W.; Zhang, Y.; Zhao, C.; Zhao, Y.; Zhou, S.; Zhou, S.; Zhu, Q.; and Zou, Y. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. arXiv:2401.02954.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Eisape, T.; Tessler, M.; Dasgupta, I.; Sha, F.; van Steenkiste, S.; and Linzen, T. 2024. A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models. arXiv:2311.00445.

Evans, J. S. B. T., ed. 1990. *Bias in Human Reasoning: Causes and Consequences*. Psychology Press.

Evans, J. S. B. T.; Barston, J. L.; and Pollard, P. 1983. On the Conflict Between Logic and Belief in Syllogistic Reasoning. *Memory & Cognition*, 11(3): 295–306.

Goel, V.; Gold, B.; Kapur, S.; and Houle, S. 1998. Neuroanatomical correlates of human reasoning. *Journal of Cognitive Neuroscience*, 12(5): 293–310.

Google DeepMind. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. Official announcement.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Roziere, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Wyatt, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; Garcia-Olano, D.; Perino, D.; Hupkes, D.; Lakomkin, E.; AlBadawy, E.; Lobanova, E.; Dinan, E.; Smith, E. M.; Radenovic, F.; Guzmán, F.; Zhang, F.; Synnaeve, G.; Lee, G.; Anderson, G. L.; Thattai, G.; Nail, G.; Mialon, G.; Pang, G.; Cucurell, G.; Nguyen, H.; Korevaar, H.; Xu, H.; Touvron, H.; Zarov, I.; Ibarra, I. A.; Kloumann, I.; Misra, I.; Evtimov, I.; Zhang, J.; Copet, J.; Lee, J.; Geffert, J.; Vranes, J.; Park, J.; Mahadeokar, J.; Shah, J.; van der Linde, J.; Billock, J.; Hong, J.; Lee, J.; Fu, J.; Chi, J.; Huang, J.; Liu, J.; Wang, J.; Yu, J.; Bitton, J.; Spisak, J.; Park, J.; Rocca, J.; Johnstun, J.; Saxe, J.; Jia, J.; Alwala, K. V.; Prasad, K.; Upasani, K.; Plawiak, K.; Li, K.; Heafield, K.; Stone, K.; El-Arini, K.; Iyer, K.; Malik, K.; Chiu, K.; Bhalla, K.; Lakhotia, K.; Rantala-Yeary, L.; van der Maaten, L.; Chen, L.; Tan, L.; Jenkins, L.; Martin, L.; Madaan, L.; Malo, L.; Blecher, L.; Landzaat, L.; de Oliveira, L.; Muzzi, M.; Pasupuleti, M.; Singh, M.; Paluri, M.; Kardas, M.; Tsimpoukelli, M.; Oldham, M.; Rita, M.; Pavlova, M.; Kambadur, M.; Lewis, M.; Si, M.; Singh, M. K.; Hassan, M.; Goyal, N.; Torabi, N.; Bashlykov, N.; Bogoychev, N.; Chatterji, N.; Zhang, N.; Duchenne, O.; Çelebi, O.; Alrassy, P.; Zhang, P.; Li, P.; Vasic, P.; Weng, P.; Bhargava, P.; Dubal, P.; Krishnan, P.; Koura, P. S.; Xu, P.; He, Q.; Dong, Q.; Srinivasan, R.; Ganapathy, R.; Calderer, R.; Cabral, R. S.; Stojnic, R.; Raileanu, R.; Maheswari, R.; Girdhar, R.; Patel, R.; Sauvestre, R.; Polidoro, R.; Sumbaly, R.; Taylor, R.; Silva, R.; Hou, R.; Wang, R.; Hosseini, S.; Chennabasappa, S.; Singh, S.; Bell, S.; Kim, S. S.; Edunov, S.; Nie, S.; Narang, S.; Raparthy, S.; Shen, S.; Wan, S.; Bhosale, S.; Zhang, S.; Vandenhende, S.; Batra, S.; Whitman, S.; Sootla, S.; Collot, S.; Gururangan, S.; Borodinsky, S.; Herman, T.; Fowler, T.; Sheasha, T.; Georgiou, T.; Scialom, T.; Speckbacher, T.; Mihaylov, T.; Xiao, T.; Karn, U.; Goswami, V.; Gupta, V.; Ramanathan, V.; Kerkez, V.; Gonguet, V.; Do, V.; Vogeti, V.; Albiero, V.; Petrovic, V.; Chu, W.; Xiong, W.; Fu, W.; Meers, W.; Martinet, X.; Wang, X.; Wang, X.; Tan, X. E.; Xia, X.; Xie, X.; Jia, X.; Wang, X.; Goldschlag, Y.; Gaur, Y.; Babaei, Y.; Wen, Y.; Song, Y.; Zhang, Y.; Li, Y.; Mao, Y.; Coudert, Z. D.; Yan, Z.; Chen, Z.; Papakipos, Z.; Singh, A.; Srivastava, A.; Jain, A.; Kelsey, A.; Shajnfeld, A.; Gangidi, A.; Victoria, A.; Goldstand, A.; Menon, A.; Sharma, A.; Boesenberg, A.; Baevski, A.; Feinstein, A.; Kallet, A.; Sangani, A.; Teo, A.; Yunus, A.; Lupu, A.; Alvarado, A.; Caples, A.; Gu, A.; Ho, A.; Poulton, A.; Ryan, A.; Ramchandani, A.; Dong, A.; Franco, A.; Goyal, A.; Saraf, A.; Chowdhury, A.; Gabriel, A.; Bharambe, A.; Eisenman, A.; Yazdan, A.; James, B.; Maurer, B.; Leonhardi, B.; Huang, B.; Loyd, B.; Paola, B. D.; Paranjape, B.; Liu, B.; Wu, B.; Ni, B.; Hancock, B.; Wasti, B.; Spence, B.; Stojkovic, B.; Gamido, B.;

Montalvo, B.; Parker, C.; Burton, C.; Mejia, C.; Liu, C.; Wang, C.; Kim, C.; Zhou, C.; Hu, C.; Chu, C.-H.; Cai, C.; Tindal, C.; Feichtenhofer, C.; Gao, C.; Civin, D.; Beaty, D.; Kreymer, D.; Li, D.; Adkins, D.; Xu, D.; Testuggine, D.; David, D.; Parikh, D.; Liskovich, D.; Foss, D.; Wang, D.; Le, D.; Holland, D.; Dowling, E.; Jamil, E.; Montgomery, E.; Presani, E.; Hahn, E.; Wood, E.; Le, E.-T.; Brinkman, E.; Arcaute, E.; Dunbar, E.; Smothers, E.; Sun, F.; Kreuk, F.; Tian, F.; Kokkinos, F.; Ozgenel, F.; Caggioni, F.; Kanayet, F.; Seide, F.; Florez, G. M.; Schwarz, G.; Badeer, G.; Swee, G.; Halpern, G.; Herman, G.; Sizov, G.; Guangyi; Zhang; Lakshminarayanan, G.; Inan, H.; Shojanazeri, H.; Zou, H.; Wang, H.; Zha, H.; Habeeb, H.; Rudolph, H.; Suk, H.; Aspegren, H.; Goldman, H.; Zhan, H.; Damlaj, I.; Molybog, I.; Tufanov, I.; Leontiadis, I.; Veliche, I.-E.; Gat, I.; Weissman, J.; Geboski, J.; Kohli, J.; Lam, J.; Asher, J.; Gaya, J.-B.; Marcus, J.; Tang, J.; Chan, J.; Zhen, J.; Reizenstein, J.; Teboul, J.; Zhong, J.; Jin, J.; Yang, J.; Cummings, J.; Carvill, J.; Shepard, J.; McPhie, J.; Torres, J.; Ginsburg, J.; Wang, J.; Wu, K.; U, K. H.; Saxena, K.; Khandelwal, K.; Zand, K.; Matosich, K.; Veeraraghavan, K.; Michelena, K.; Li, K.; Jagadeesh, K.; Huang, K.; Chawla, K.; Huang, K.; Chen, L.; Garg, L.; A, L.; Silva, L.; Bell, L.; Zhang, L.; Guo, L.; Yu, L.; Moshkovich, L.; Wehrstedt, L.; Khabsa, M.; Avalani, M.; Bhatt, M.; Mankus, M.; Hasson, M.; Lennie, M.; Reso, M.; Groshev, M.; Naumov, M.; Lathi, M.; Keneally, M.; Liu, M.; Seltzer, M. L.; Valko, M.; Restrepo, M.; Patel, M.; Vyatskov, M.; Samvelyan, M.; Clark, M.; Macey, M.; Wang, M.; Hermoso, M. J.; Metanat, M.; Rastegari, M.; Bansal, M.; Santhanam, N.; Parks, N.; White, N.; Bawa, N.; Singhal, N.; Egebo, N.; Usunier, N.; Mehta, N.; Laptev, N. P.; Dong, N.; Cheng, N.; Chernoguz, O.; Hart, O.; Salpekar, O.; Kalinli, O.; Kent, P.; Parekh, P.; Saab, P.; Balaji, P.; Rittner, P.; Bontrager, P.; Roux, P.; Dollar, P.; Zvyagina, P.; Ratanchandani, P.; Yuvraj, P.; Liang, Q.; Alao, R.; Rodriguez, R.; Ayub, R.; Murthy, R.; Nayani, R.; Mitra, R.; Parthasarathy, R.; Li, R.; Hogan, R.; Battey, R.; Wang, R.; Howes, R.; Rinott, R.; Mehta, S.; Siby, S.; Bondu, S. J.; Datta, S.; Chugh, S.; Hunt, S.; Dhillon, S.; Sidorov, S.; Pan, S.; Mahajan, S.; Verma, S.; Yamamoto, S.; Ramaswamy, S.; Lindsay, S.; Lindsay, S.; Feng, S.; Lin, S.; Zha, S. C.; Patil, S.; Shankar, S.; Zhang, S.; Zhang, S.; Wang, S.; Agarwal, S.; Sajuyigbe, S.; Chintala, S.; Max, S.; Chen, S.; Kehoe, S.; Satterfield, S.; Govindaprasad, S.; Gupta, S.; Deng, S.; Cho, S.; Virk, S.; Subramanian, S.; Choudhury, S.; Goldman, S.; Remez, T.; Glaser, T.; Best, T.; Koehler, T.; Robinson, T.; Li, T.; Zhang, T.; Matthews, T.; Chou, T.; Shaked, T.; Vontimitta, V.; Ajayi, V.; Montanez, V.; Mohan, V.; Kumar, V. S.; Mangla, V.; Ionescu, V.; Poenaru, V.; Mihailescu, V. T.; Ivanov, V.; Li, W.; Wang, W.; Jiang, W.; Bouaziz, W.; Constable, W.; Tang, X.; Wu, X.; Wang, X.; Wu, X.; Gao, X.; Kleinman, Y.; Chen, Y.; Hu, Y.; Jia, Y.; Qi, Y.; Li, Y.; Zhang, Y.; Zhang, Y.; Adi, Y.; Nam, Y.; Yu; Wang; Zhao, Y.; Hao, Y.; Qian, Y.; Li, Y.; He, Y.; Rait, Z.; DeVito, Z.; Rosnbrick, Z.; Wen, Z.; Yang, Z.; Zhao, Z.; and Ma, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Jenkinson, A. J. 1985. Prior Analytics. In Barnes, J., ed., *Complete Works of Aristotle, Volume 1: The Revised Oxford Translation*, 39–113. Princeton University Press.

Jiang, B.; Xie, Y.; Hao, Z.; Wang, X.; Mallick, T.; Su, W. J.; Taylor, C. J.; and Roth, D. 2024. A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners. arXiv:2406.11050.

Khemlani, S.; and Johnson-Laird, P. N. 2012. Theories of the Syllogism: A Meta-Analysis. *Psychological Bulletin*, 138(3): 427–457.

Lappin, S.; and Fox, C., eds. 2015. *The Handbook of Contemporary Semantic Theory*. Chichester, West Sussex: Wiley.

Liu, Y.; Guo, Z.; Liang, T.; Shareghi, E.; Vulić, I.; and Collier, N. 2025. Aligning with Logic: Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models. arXiv:2410.02205.

Miller, G. A. 1992. WordNet: A Lexical Database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Mondorf, P.; and Plank, B. 2024. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey. arXiv:2404.01869.

Morishita, T.; Morio, G.; Yamaguchi, A.; and Sogawa, Y. 2024. Enhancing Reasoning Capabilities of LLMs via Principled Synthetic Logic Corpus. arXiv:2411.12498.

OpenAI. 2024. GPT-4o System Card. OpenAI Technical Report.

Parmar, M.; Patel, N.; Varshney, N.; Nakamura, M.; Luo, M.; Mashetty, S.; Mitra, A.; and Baral, C. 2024. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models. arXiv:2404.15522.

Sánchez, V. C. 1991. *Studies on Natural Logic and Categorial Grammar*. Ph.D. thesis, University of Amsterdam.

Stenning, K.; and van Lambalgen, M. 2008. *Human Reasoning and Cognitive Science*. Boston, USA: MIT Press.

van Benthem, J. 1983. Determiners and Logic. *Linguistics and Philosophy*, 6(4): 447–478.

Woodworth, R. S.; and Sells, S. B. 1935. An Atmosphere Effect in Formal Syllogistic Reasoning. *Journal of Experimental Psychology*, 18(4): 451.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.