GigitAI at SemEval-2026 Task 11: Hybrid Symbolic-Neural Approach for Syllogistic Validity Classification

Saran Krishnasamy*
GigitAI
San Francisco, CA
saran@gigit.ai

Abstract

We present our system for SemEval-2026 Task 11 on classifying whether syllogisms are logically valid. The main challenge is that language models tend to judge arguments based on whether the conclusion sounds true in the real world, rather than whether it follows logically from the premises. We evaluate direct prompting across six models (GPT-4o, GPT-5.2, o3, o3-mini, Claude Opus 4.6, Claude Sonnet 4) with three prompt strategies, finding that even the best achieves only 89.5% accuracy. Our best-performing system splits the task into two parts: GPT-4o-mini extracts the logical structure, then deterministic rules check validity, enhanced with bidirectional premise checking, predicate negation post-processing, and a targeted rule-based fallback for double negation. This achieves 98.95% accuracy on Subtask 1 (combined score 57.74) and 85.8% validity accuracy on Subtask 2. We also explore self-consistency with symbolic verification (93.1%), content abstraction, activation steering, contrastive fine-tuning, RLVR, and diffusion-based reasoning, finding that content abstraction surprisingly degrades performance, revealing that semantic content provides essential parsing scaffolding alongside the bias it introduces.

1 Introduction

The ability to distinguish valid from invalid arguments, independent of whether the conclusion happens to be true, lies at the heart of logical reasoning. Consider these two syllogisms:

> *All dogs are mammals. All mammals breathe. Therefore, all dogs breathe.*
>
> *All fish are birds. All birds can fly. Therefore, all fish can fly.*

Both arguments share identical logical structure, the classic "Barbara" form that has been recognized as valid since Aristotle's *Prior Analytics*. Yet humans systematically judge the first as "more valid" than the second, a phenomenon known as *belief bias* (Evans et al., 1983). When the conclusion aligns with our knowledge of the world, we accept the argument more readily; when it contradicts reality, we become skeptical of the reasoning itself.

This conflation of logical validity with factual truth poses a fundamental challenge for language models trained on natural text. SemEval-2026 Task 11 (Valentino et al., 2026) directly probes this capability, joining other benchmarks that test logical reasoning in language models (Liu et al., 2020). The task asks systems to classify syllogistic validity while remaining robust to content plausibility, and introduces the Total Content Effect metric to quantify how much a system's judgments are swayed by whether conclusions "sound true."

Our approach stems from a simple observation: while large language models struggle to *judge* logical validity directly, they excel at *parsing* natural language into structured representations. Rather than asking GPT-4o-mini "Is this syllogism valid?", a question that triggers belief bias, we ask it to extract the logical form: "What is the quantifier? What are the subject and predicate terms?" We then validate this extracted structure against the 24 canonical syllogistic forms using deterministic rules that are, by construction, immune to content effects.

2 Background

SemEval-2026 Task 11 (Valentino et al., 2026) evaluates syllogistic reasoning across four subtasks; we participate in the two English-language subtasks (Subtasks 3 and 4 are multilingual equivalents). Subtask 1 requires classifying whether a two-premise syllogism is logically valid. Subtask 2 presents 5–8 premises, only two of which are relevant; systems must identify the relevant pair

*Code, prompts, and outputs available at https://github.com/sarankrish/semeval2026-task11

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3432–3441

and judge validity. The evaluation uses the Total Content Effect (TCE) to measure how much a system's accuracy varies with content plausibility, combined with accuracy into a single score: $\text{accuracy}/(1 + \ln(1 + \text{TCE}))$.

Each proposition in a categorical syllogism takes one of four forms: universal affirmative (A: "All S are P"), universal negative (E: "No S are P"), particular affirmative (I: "Some S are P"), or particular negative (O: "Some S are not P"). Combined with four possible *figures* (middle term positions), this yields 256 possible forms, of which exactly 24 are valid (Corcoran, 1972).

Belief bias, the tendency to judge arguments based on conclusion plausibility rather than logical structure, is well-documented in humans (Evans et al., 1983; Khemlani and Johnson-Laird, 2012) and has been shown to affect LLMs similarly (Dasgupta et al., 2024). This motivates neural-symbolic approaches (d'Avila Garcez et al., 2019; Huang and Chang, 2023) that separate language understanding from logical judgment.

Related work. Three lines of work motivate our approach. *Hybrid neuro-symbolic systems for logical reasoning*: LINC (Olausson et al., 2023) translates premises to first-order logic for theorem provers; Logic-LM (Pan et al., 2023) self-refines symbolic formulations using solver feedback; SatLM (Ye et al., 2023) uses declarative specifications with SMT solvers. We apply this paradigm specifically to syllogisms, where the discrete 24-form target enables a particularly tight neural-symbolic interface. *Syllogistic reasoning in LLMs*: Eisape et al. (2024) document figural and content effects in PaLM 2; Bertolazzi et al. (2024) systematically study chain-of-thought, in-context learning, and supervised fine-tuning, finding SFT on pseudo-word syllogisms most effective at mitigating content bias; Seals and Shalin (2024) similarly emphasize biases in LLM deductive reasoning. *Activation-level mitigation*: Maraia et al. (2026) learn Abstractor networks targeting an explicit abstract reasoning space, a more sophisticated alternative to the difference-of-means probe we explore in §4.4.

3 System Description

Figure 1 shows our system architecture for both subtasks.

Model	Best Prompt	Acc.	TCE
Claude Opus 4.6	Debiased	89.5%	1.44
o3-mini	Debiased	85.8%	5.60
GPT-5.2	Debiased	85.3%	1.54
GPT-4o	Debiased	84.2%	24.24
Claude Sonnet 4	Debiased	80.1%	7.21
o3	Debiased	78.5%	3.16

Table 1: Best direct prompting result per model on the 960-example training set. All models perform best with the debiased prompt. See Appendix C for all 18 configurations.

3.1 The Failure of Direct Prompting

Our initial experiments revealed the severity of belief bias across a wide range of current LLMs. We evaluated six models–GPT-4o, GPT-5.2, o3, o3-mini, Claude Opus 4.6, and Claude Sonnet 4–with three prompting strategies (structure-focused, debiased, and few-shot) on the full 960-example training set. Table 1 summarizes the best result per model (see Appendix C for all 18 configurations).

Several patterns emerge.[1] First, the *debiased* prompt, which explicitly warns that premises may be factually false and provides contrastive examples, is the best strategy for every model. Second, prompt sensitivity can be extreme: Claude Opus 4.6 achieves the highest accuracy (89.5%) with the debiased prompt, yet drops to near chance (49.9%) with few-shot or structure-focused prompts, where it generates reasoning chains that exceed the output token limit and default to "valid."[2] Third, deeper reasoning can amplify bias: o3, despite being a stronger reasoning model than GPT-4o, achieves lower accuracy (78.5% vs. 84.2%), suggesting its extended chain-of-thought reasoning (Wei et al., 2022) gives it more opportunity to incorporate world knowledge, leading it to reject implausible-but-valid syllogisms. Even the best direct prompting result (89.5%) falls short of formal methods, motivating our two-stage approach.

[1] All API calls used temperature 0 and default maximum output tokens via the official OpenAI and Anthropic APIs. The Claude few-shot and structure-focused configurations exceeded this limit during reasoning generation; truncated outputs lacking a final "VALID"/"INVALID" token defaulted to "VALID" in our parsing.

[2] Our few-shot prompt instructs abstract-form reasoning ("convert to abstract form (A, B, C)") before requiring a single-word output, a contradiction that may partly explain few-shot underperformance; the cleaner debiased prompt avoids this issue. Additionally, the few-shot examples use abstract variables while inputs are concrete natural language, and the prompt does not explicitly bridge this gap.

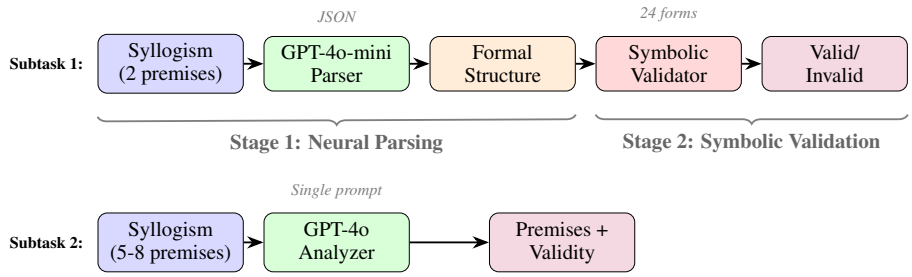

Figure 1: System architecture. Subtask 1 uses a two-stage approach: Stage 1 (neural parsing) uses GPT-4o-mini to extract formal structure, then Stage 2 (symbolic validation) checks against the 24 valid syllogistic forms. Subtask 2 uses a single GPT-4o prompt to jointly identify relevant premises and assess validity.

3.2 Stage 1: Neural Parsing

The first stage uses GPT-4o-mini (OpenAI, 2023) to parse natural language syllogisms into formal structures. Given a syllogism like *"Every dog is a mammal. All mammals are animals. Therefore, every dog is an animal,"* the model extracts:

	Quantifier	**Subject**	**Predicate**
Premise 1	All	dog	mammal
Premise 2	All	mammal	animal
Conclusion	All	dog	animal

The prompt instructs the model to output structured JSON, focusing purely on extraction rather than evaluation (Liu et al., 2023). This reframing appears to largely bypass belief bias, since the model isn't being asked whether the argument is good, just what its components are.

Predicate negation post-processing. We found that GPT-4o-mini sometimes returns negation within the predicate (e.g., quantifier "All", predicate "not integers") rather than correctly flipping the quantifier. We apply a mechanical post-processing step: if the predicate begins with "not", we strip it and flip the quantifier (All ↔ No, Some ↔ Some…not). This corrects parses like *"Every X is not a Y"* from All(X, not-Y) to the correct No(X, Y).

3.3 Stage 2: Symbolic Validation

Given the parsed structure, validation becomes mechanical. We convert each proposition to its categorical type (A, E, I, or O), identify the figure based on middle term position, and check whether the resulting combination appears in our list of the 24 canonical valid forms (Corcoran, 1972), which include both the 15 unconditionally valid moods and 9 conditionally valid moods that derive particular conclusions from universal premises (e.g., Barbari, Darapti). This stage is entirely deterministic and content-blind: *"All zorbs are flimps"* receives identical treatment to *"All dogs are mammals."*

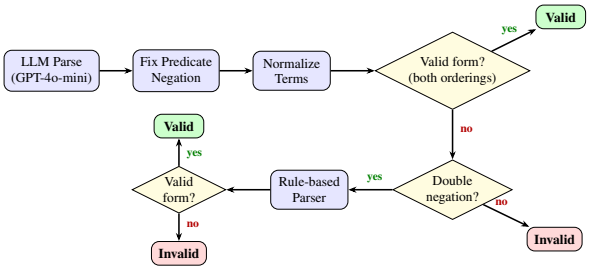

Figure 2: Validation pipeline for Subtask 1. After LLM parsing and term normalization, bidirectional premise checking tests both orderings against the 24 valid forms. If invalid and double negation is detected, a targeted rule-based fallback re-parses the syllogism.

Figure 2 shows the full validation pipeline. Three enhancements improve accuracy beyond naive form lookup: (1) *fuzzy term matching* normalizes morphological variation by stripping articles, quantifiers, and descriptive prefixes (e.g., *"things that are dogs"* → *"dogs"*) and comparing with singular/plural tolerance; (2) *bidirectional premise checking* tries both assignments of premises as major and minor, with term consistency verification to prevent false positives; and (3) a *targeted rule-based fallback* for double negation, where GPT-4o-mini consistently misparses *"There are no X that are not Y"* as No(X, Y) rather than the correct All(X, Y). The fallback is narrow enough to avoid the high false-positive rate of full rule-based parsing (73% accuracy) while rescuing these specific cases.

3.4 Premise Retrieval for Subtask 2

Subtask 2 presented an additional challenge: identifying which two of 5–8 premises are relevant to the conclusion. Our initial approach attempted structured term-matching: parse all premises and find the pair sharing terms with the conclusion.

This achieved only 22.9% F1 on the training set, failing on synonyms, paraphrases, and complex noun phrases. In this version, validity was assigned heuristically: when two premises sharing terms with the conclusion (via a middle term) could be located and formed a valid syllogism, the example was labeled valid; otherwise it defaulted to invalid. This default-to-invalid behavior produced a heavy bias (84.4% predicted invalid), accounting for the low validity accuracy (58.3%).

The solution was to leverage LLM semantic understanding more directly. Rather than parsing all premises independently, we prompt GPT-4o to jointly identify the relevant premises and assess validity in a single pass. The model is instructed to find the two premises that share a "middle term" connecting to the conclusion.

4 Additional Approaches Explored

4.1 Content Abstraction

A natural hypothesis for reducing belief bias is to remove semantic content entirely. We replaced content terms with nonsense words (e.g., *"All zorbs are flimps"*). On Subtask 1, accuracy dropped sharply: rule-based fell from 73.0% to 54.9%, LLM-parsed from 93.2% to 59.6%, with extreme bias toward predicting invalid (∼20% valid accuracy, ∼99% invalid). On Subtask 2, validity accuracy dropped from 85.8% to 60.9%, premise F1 from 86.3% to 31.8%, and TCE *increased* from 11.2 to 31.1. The parsing prompt itself was unchanged for these experiments; we hypothesize that the sharp degradation reflects, in part, reduced parsing reliability on nonsense vocabulary, consistent with the parallel drop in both rule-based (73.0% to 54.9%) and LLM-parsed (93.2% to 59.6%) accuracy. Semantic content appears to serve a dual role: introducing bias but also providing scaffolding for correct parsing (Gendron et al., 2024).

4.2 Z3 SMT Solver Validation

We implemented a formal verification backend using the Z3 SMT solver (De Moura and Bjørner, 2008). Each syllogism is encoded in first-order logic, and validity is checked by testing whether (premises $\wedge$ $\neg$conclusion) is unsatisfiable. Z3 agrees with our 24-form lookup on all standard forms, confirming correctness. As a standalone system with rule-based parsing, Z3 achieves 75.62% on the training set (97% on invalid, 55% on valid), confirming that the bottleneck is parsing rather than validation logic.

4.3 Self-Consistency with Symbolic Verification

Inspired by self-consistency prompting (Wang et al., 2023), we sample $N = 5$ reasoning paths from GPT-4o-mini at temperature 0.7, parse each into formal structure, validate with the symbolic validator, and use majority voting among verified paths. This achieves 93.1% accuracy (TCE 3.08, combined 38.7) on the training set, outperforming direct prompting (89.5%) by combining LLM diversity with symbolic guarantees.

4.4 Activation Steering

Drawing on representation engineering (Zou et al., 2023) and contrast-consistent search (Burns et al., 2022), we explored whether transformer models encode logical validity as a linear direction in activation space. We compute a *validity direction* from DeBERTa-v3 (He et al., 2023) activations as the difference of class means, classifying inputs by projection with multi-layer ensemble voting. This achieves 64.1% accuracy (TCE 6.32, combined 21.4) on the training set, suggesting pre-trained representations encode only a weak validity signal. Concurrent work by Maraia et al. (2026) takes a more sophisticated approach, learning lightweight Abstractor networks that map content-conditioned residual states to an abstract reasoning space defined from paired content-laden and abstract syllogisms; their multi-layer interventions during the forward pass yield more reliable content-effect mitigation than the single difference-of-means probe explored here.

4.5 RLVR: Reinforcement Learning with Verifiable Rewards

Our symbolic validator provides a perfect reward signal for RL (Lightman et al., 2023). We implemented GRPO (Shao et al., 2024), as used in DeepSeek-R1 (Guo et al., 2025), to train Qwen2.5-3B (Qwen et al., 2025) with rewards of $+1.0$ (correct), -1.0 (wrong), and 0.0 (unparseable), using 4-bit quantization with LoRA (Hu et al., 2022).

A key challenge was the cold-start problem: the base model could not produce parseable output (e.g., *"ANSWER: TRUE"*), so all responses received zero reward and no gradient signal flowed. Following the supervised fine-tuning (SFT) warmup strategy used in InstructGPT (Ouyang et al., 2022) and DeepSeek-R1, we first train the

model for one epoch on ground-truth input-output pairs to teach the expected format. After this warmup, 100% of outputs are parseable and RL training proceeds normally. After one RL epoch, accuracy reaches 57.8% (up from 52% at initialization), with average reward steadily increasing. The modest gains likely reflect the limited capacity of a 3B-parameter model for this task.

4.6 Contrastive Fine-tuning

Inspired by contrastive embeddings (Gao et al., 2021), we fine-tuned DeBERTa-v3 with combined classification and contrastive loss, where positive pairs share validity regardless of plausibility and hard negatives share plausibility but differ in validity. We use balanced quadrant sampling and an adversarial plausibility head with gradient reversal. On a stratified 15% held-out validation split (144 examples), this achieves 76.2% accuracy (TCE 9.72, combined 22.6).

4.7 Diffusion-based Reasoning

We explored discrete diffusion (Austin et al., 2021) to iteratively refine logical structure. The model encodes syllogisms into a structured latent space, adds noise, and trains a denoiser to recover the valid form. On the training set, accuracy reaches 99.1% (TCE 0.42, combined 73.4) after 10 epochs, climbing from 74.8% at epoch 1. However, since evaluation uses training data, this likely reflects overfitting rather than genuine reasoning ability; the high combined score is not comparable to systems evaluated on held-out data.

5 Results and Analysis

5.1 Subtask 1: Classification

Table 2 summarizes our Subtask 1 results across all approaches. The best direct prompting (Claude Opus 4.6 debiased, 89.5%) far exceeds naive prompting but falls short of symbolic methods. Our final hybrid system achieves 98.95% accuracy on the 191-example test set with a TCE of 1.04, yielding a combined score of 57.74. On the 960-example training set, the final enhanced-parsing system achieves 93.2% (65 errors: 42 FN, 23 FP), lower than its test performance because the larger training set contains more parsing edge cases (adversarial negation, compound terms) that were rarer in the smaller test set; targeted enhancements such as the double-negation fallback were added in response to specific parsing failures observed during development.[3]

Approach	Val. Acc.	TCE	Comb.
Direct prompting (best)	89.5%	1.44	47.3
Direct prompting (worst)	49.9%	0.42	37.0
Rule-based only	73.0%	26.7	16.9
Content abstraction	59.6%	1.71	29.9
Activation steering	64.1%	6.32	21.4
Contrastive fine-tuning	76.2%*	9.72	22.6
Self-consistency (N=5)	93.1%	3.08	38.7
Z3 + rule-based parsing	75.6%	3.17	31.2
Symbolic validation (v1)	94.24%	2.08	44.33
+ Enhanced parsing (v2)	**98.95%**	**1.04**	**57.74**

Table 2: Subtask 1 results. All rows except the last two are evaluated on the training set; the last two are official test set submissions. *Evaluated on held-out validation split. See Appendix C for full direct prompting breakdown.

The 2 remaining errors are both false negatives (valid syllogisms predicted as invalid), with zero false positives, one on plausible content and one on implausible, showing no content bias at the error level:

> *There are no animals that are non-aquatic and also fish. A portion of the animals that are mammals are also non-aquatic. It must be the case that some mammals are not fish.*
>
> **Gold: Valid** **Predicted: Invalid** (Plausible)

> *Every single item that is a number is not an integer. Of the items that are prime numbers, some are integers. It is a correct deduction that some prime numbers are not numbers.*
>
> **Gold: Valid** **Predicted: Invalid** (Implausible)

The first contains a compound term with embedded negation (*"non-aquatic and also fish"*) that confuses parsing. The second has an unusual negation pattern (*"is a number is not an integer"*) where GPT-4o-mini non-deterministically misparses the quantifier. Despite this balance, the official TCE of 1.04 indicates some residual content sensitivity, likely arising from subtle parsing variation with content plausibility.

5.2 Subtask 2: Retrieval and Classification

Table 3 presents our Subtask 2 results across three system versions. The dramatic improvement from term matching (v1) to direct LLM analysis (v2) highlights the importance of leveraging semantic understanding for premise identification.

[3]The earlier symbolic system (without enhanced parsing) achieves 93.72% on the test set; training-set performance for this version was not separately tracked.

Approach	Comb.	Val.	F1	TCE
v1: Term matching	8.79	58.3%	22.9%	36.4
v2: LLM direct	**24.6**	**85.8%**	**86.3%**	**11.2**
v3: +Abstraction	10.38	60.9%	31.8%	31.1

Table 3: Subtask 2 results. TCE = Total Content Effect (lower is better). Direct LLM analysis dramatically outperforms structured term matching.

Our best system (v2) achieved 85.79% validity accuracy and 86.32% premise retrieval F1, for a combined score of 24.6 (rank 13 of 29). The joint accuracy, getting both premises and validity correct, was 78.42%.

Error analysis. The 27 validity errors show a strong bias toward predicting VALID: 21 false positives versus only 6 false negatives. Of the false positives, 11 occurred on plausible-but-invalid syllogisms, indicating residual belief bias. The dominant pattern is the *undistributed middle fallacy*:

> *All doctors are professionals. All lawyers are professionals.*
> *Therefore, some lawyers are not doctors.*
> **Gold: Invalid Predicted: Valid**

The middle term "professionals" appears only as the predicate of universal affirmative premises and is never distributed, so no valid conclusion follows. Yet the conclusion is obviously true in the real world, creating a content-bias trap: the model accepts the argument because the conclusion *sounds right*, not because it *follows logically*.

The 20 premise selection errors show a different pattern: the model identifies premises with semantic similarity rather than logical relevance:

> *[0] All spiders have exoskeletons. [1] Anything that is an insect has an exoskeleton. . . . [5] Any creature that is an ant is an insect.*
> *Conclusion: All ants have exoskeletons.*
> **Gold: [1, 5] Predicted: [0, 5]**

The model correctly identified that ants are insects (premise 5), but selected "all spiders have exoskeletons" instead of "all insects have exoskeletons," drawn to a semantically related arthropod rather than the logically necessary general claim.

6 Conclusion

We presented a hybrid symbolic-neural system that separates LLM-based parsing from deterministic symbolic validation, achieving 98.95% accuracy on Subtask 1 (combined 57.74). Our evaluation across six models confirms that belief bias affects all current LLMs, with the counterintuitive finding that stronger reasoning models (o3) perform worse than standard models (GPT-4o), that prompt choice can swing accuracy by 40 percentage points (Claude Opus 4.6: 49.9% to 89.5%), and that content abstraction degrades rather than improves performance.

Limitations. Our system is intentionally specialized to categorical syllogisms; the closed structure of the 24 valid Aristotelian forms enables a particularly tight neural-symbolic interface, but the validation logic does not transfer directly to broader logical reasoning tasks. Extension to first-order logic (as in LINC; Olausson et al. 2023) would require swapping our 24-form lookup for a general theorem prover. Additionally, the system depends on closed-source GPT-4o-mini for parsing, limiting full reproducibility; replacing it with an open-source parser is left to future work.

Future work. Four directions are promising: (1) scaling RLVR to larger models (8B+), where the 3B model's limited capacity appears to bottleneck learning despite a clear training signal; (2) extending the two-stage philosophy to Subtask 2, using LLMs for premise selection but symbolic validation for validity; (3) improving neural parsing to eliminate the last two errors, potentially through ensemble parsing (multiple GPT-4o-mini samples with majority voting on the parsed structure) or specialized fine-tuning on synthetic data targeting compound terms and embedded negation, the two specific patterns observed in our failure cases; and (4) exploring whether prompting the LLM to convert syllogisms directly to canonical (A/E/I/O) form, rather than extracting individual quantifier-subject-predicate triples, would be more robust to negation and compound-term parsing failures.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Pro-*

cessing, pages 13882–13905, Miami, Florida, USA. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *ArXiV*.

John Corcoran. 1972. Completeness of an ancient logic. *Journal of Symbolic Logic*, 37(4):696–702.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *Preprint*, arXiv:2207.07051.

Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Preprint*, arXiv:1905.06088.

Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: an efficient smt solver. In *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, TACAS'08/ETAPS'08, page 337–340, Berlin, Heidelberg. Springer-Verlag.

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

J. St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Sangeet Khemlani and P N Johnson-Laird. 2012. Theories of the syllogism: A meta-analysis. *Psychol Bull*, 138(3):427–457.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization. Main track.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Gabriele Maraia, Marco Valentino, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2026. Abstract activation spaces for content-invariant reasoning in large language models. *Preprint*, arXiv:2602.02462.

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

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with

human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

S Seals and Valerie Shalin. 2024. Evaluating the deductive competence of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8614–8630, Mexico City, Mexico. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026. Semeval-2026 task 11: Disentangling content and formal reasoning in large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. Satlm: satisfiability-aided language models using declarative prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A System Configuration

Our Subtask 1 system uses GPT-4o-mini (`gpt-4o-mini`) with temperature 0 for deterministic parsing. The symbolic validator checks against the 24 canonical valid syllogistic forms (15 unconditionally valid moods plus 9 conditionally valid moods). Post-processing includes predicate negation correction, bidirectional premise checking with term consistency verification, and targeted rule-based fallback for double negation patterns. For Subtask 2, we use GPT-4o (`gpt-4o-2024-08-06`) with temperature 0 for joint premise identification and validity assessment. For content abstraction experiments, we used a fixed vocabulary of 20 nonsense words (zorbs, flimps, glerks, etc.) with case-preserving, longest-first replacement.

B Prompts

B.1 Subtask 1: Parsing Prompt

The following prompt is sent to GPT-4o-mini to extract formal logical structure from natural language syllogisms:

```
Parse this syllogism into formal logical
structure.

A syllogism has exactly 3 parts:
1.    Major premise (first premise,
contains the predicate of the
conclusion)
2.    Minor premise (second premise,
contains the subject of the conclusion)
3. Conclusion (starts with "therefore",
"thus", "so", "hence", "consequently",
etc.)

For each proposition, identify:
- Quantifier: "all/every/each" = ALL,
"no/none" = NO, "some/at least one" =
SOME, "some...not/not all" = SOME_NOT
-   Subject:    the   first   term   (a
category/class noun phrase)
-  Predicate:   the   second   term   (a
category/class noun phrase)

Normalize terms to their simplest form
(e.g., "things that are dogs" -> "dogs",
"creatures with feathers" -> "feathered
creatures").

SYLLOGISM:
{syllogism}
```

```
Respond in this exact JSON format:
{
  "major_premise": {"quantifier":
"ALL|NO|SOME|SOME_NOT", "subject":
"term1", "predicate": "term2"},
  "minor_premise": {"quantifier":
"ALL|NO|SOME|SOME_NOT", "subject":
"term1", "predicate": "term2"},
  "conclusion": {"quantifier":
"ALL|NO|SOME|SOME_NOT", "subject":
"term1", "predicate": "term2"}
}
Only output the JSON, nothing else.
```

B.2 Subtask 2: Analysis Prompt

The following prompt is sent to GPT-4o to jointly identify relevant premises and assess validity:

```
Analyze this syllogism with multiple
premises and a conclusion.

SYLLOGISM:
{syllogism}

Your task:
1. Identify the CONCLUSION (usually
after "therefore", "thus", "hence", "it
follows", etc.)
2. Find which TWO premises are logically
relevant to derive this conclusion
3. Determine if the conclusion LOGICALLY
follows from those two premises
(ignoring real-world plausibility)

A valid syllogism has the form:
- Premise 1: relates term A to term B
- Premise 2: relates term B to term C
- Conclusion: relates term A to term C

The premises share a "middle term" (B)
that connects the subject and predicate
of the conclusion.

IMPORTANT: Focus on LOGICAL structure,
not whether statements are true in the
real world. "All fish are mammals" is a
valid premise even though it's factually
false.

Respond in this exact JSON format:
{
  "conclusion_text": "the conclusion
statement",
  "relevant_premise_indices": [index1,
index2],
  "reasoning": "brief explanation",
  "validity": true/false,
  "validity_reasoning": "why the
conclusion does/doesn't logically
follow"
}

IMPORTANT: Premise indices are 0-based
(first premise is 0, second is 1, etc.).
Count premises from the beginning, NOT
including the conclusion.

Only output the JSON, nothing else.
```

B.3 Direct Prompting Strategies

The following three prompt strategies were used in the direct prompting experiments (Table 4).

Structure-focused.

```
You are a logic expert. Analyze this
syllogism's logical STRUCTURE only.
IMPORTANT: Ignore whether the statements
are true in the real world. Focus ONLY
on logical form.

Syllogism: {syllogism}

Instructions:
1. Extract the logical form by replacing
content words with variables (A, B, C)
2. Identify the quantifiers (All, Some,
No)
3. Determine if the abstract form is
logically valid

A syllogism is VALID if the conclusion
MUST be true whenever the premises are
true.
A syllogism is INVALID if the conclusion
could be false even when premises are
true.

Answer ONLY "VALID" or "INVALID":
```

Debiased.

```
CRITICAL: You must judge logical
validity INDEPENDENT of real-world
truth.

Example of a VALID syllogism with false
premises:
"All fish can fly. All salmon are fish.
Therefore, all salmon can fly."
This is VALID because IF the premises
were true, the conclusion would follow.

Example of an INVALID syllogism with
true statements:
"All dogs are mammals. All cats are
mammals. Therefore, all cats are dogs."
This is INVALID because the conclusion
doesn't follow from the premises.

Now evaluate this syllogism for LOGICAL
validity only:

Syllogism: {syllogism}

Remember: A syllogism can be logically
valid even with absurd content, and
logically invalid even with true
statements.

Answer ONLY "VALID" or "INVALID":
```

Few-shot.

```
You are a formal logic expert. Determine
if a syllogism is VALID or INVALID.

A syllogism is VALID only if the
conclusion NECESSARILY follows from the
premises.
A syllogism is INVALID if there's ANY
possible world where premises are true
but conclusion is false.

EXAMPLES:

Example 1 - VALID:
"All A are B. All B are C. Therefore,
all A are C."
Analysis: If all A are B, and all B are
C, then all A must be C. VALID.
```

```
Example 2 - INVALID:
"All A are B. All C are B. Therefore,
all C are A."
Analysis: A and C both being subsets
of B doesn't mean C is a subset of A.
INVALID.

Example 3 - INVALID:
"All A are B. No C is A. Therefore, no
C is B."
Analysis: C could still be B through
other means even if C is not A. INVALID.

Example 4 - VALID:
"No A is B. All C are A. Therefore, no
C is B."
Analysis: If no A is B and all C are A,
then no C can be B. VALID.

Example 5 - INVALID:
"Some A are B. Some B are C. Therefore,
some A are C."
Analysis: The A's that are B might not
be the same B's that are C. INVALID.

Now analyze:
Syllogism: {syllogism}

First, convert to abstract form (A, B,
C). Then check if it matches a valid
syllogistic pattern.

Answer ONLY "VALID" or "INVALID":
```

C Full Direct Prompting Results

Table 4 reports all 18 model-prompt configurations evaluated on the full 960-example training set.

Model	Prompt	Acc.	TCE
Claude Opus 4.6	Debiased	89.5%	1.44
o3-mini	Debiased	85.8%	5.60
GPT-5.2	Debiased	85.3%	1.54
GPT-4o	Debiased	84.2%	24.24
o3-mini	Few-shot	81.6%	4.53
GPT-5.2	Few-shot	81.4%	12.40
GPT-5.2	Structure	80.3%	6.76
Claude Sonnet 4	Debiased	80.1%	7.21
o3	Debiased	78.5%	3.16
o3	Structure	78.5%	3.15
o3	Few-shot	78.3%	2.31
GPT-4o	Few-shot	75.0%	75.70
GPT-4o	Structure	74.2%	58.20
o3-mini	Structure	56.5%	2.10
Claude Sonnet 4	Few-shot	50.0%	0.00
Claude Sonnet 4	Structure	50.0%	0.00
Claude Opus 4.6	Few-shot	49.9%	0.42
Claude Opus 4.6	Structure	49.9%	1.26

Table 4: Full direct prompting results on the 960-example training set. Claude Opus 4.6 shows the most extreme prompt sensitivity: 89.5% with debiased vs. 49.9% with other prompts. GPT-4o exhibits the highest TCE (24–76), while o3 shows consistently low TCE ($\sim$2–3) but also lower accuracy.

D Content Abstraction Prompt

The following prompt extracts content terms for replacement with nonsense words:

```
Extract all content terms (nouns,
noun phrases, categories) from this
syllogism.
Content terms are the subjects and
predicates - things like "dogs",
"mammals", "red things", "people who
exercise", etc.
Do NOT include logical words like
"all", "some", "no", "are", "therefore",
"hence", etc.

SYLLOGISM:
{syllogism}

Return a JSON array of unique content
terms found, in order of first
appearance.
Example: ["dogs", "mammals", "animals"]

Only output the JSON array, nothing
else.
```