FregeLogic at SemEval 2026 Task 11: A Hybrid Neuro-Symbolic Architecture for Content-Robust Syllogistic Validity Prediction

Adewale Akinfaderin
Amazon Web Services
Seattle, WA
akinfaa@amazon.com

Nafi Diallo
Amazon Web Services
Seattle, WA
nafid@amazon.com

Abstract

We present FregeLogic, a hybrid neuro-symbolic system for SemEval-2026 Task 11 (Subtask 1), which addresses syllogistic validity prediction while reducing content effects on predictions. Our approach combines an ensemble of five LLM classifiers, spanning three open-weights models (Llama 4 Maverick, Llama 4 Scout, and Qwen3-32B) paired with varied prompting strategies, with a Z3 SMT solver that serves as a formal logic tiebreaker. The central hypothesis is that LLM disagreement within the ensemble signals likely content-biased errors, where real-world believability interferes with logical judgment. By deferring to Z3's structurally-grounded formal verification on these disputed cases, our system achieves 94.3% accuracy with a content effect of 2.85 and a combined score of 41.88 in nested 5-fold cross-validation on the dataset (N=960). This represents a 2.76-point improvement in combined score over the pure ensemble (39.12), with a 0.9% accuracy gain, driven by a 16% reduction in content effect (3.39 $\rightarrow$ 2.85). Adopting structured-output API calls for Z3 extraction reduced failure rates from $\sim$22% to near zero, and an Aristotelian encoding with existence axioms was validated against task annotations. Our results suggest that targeted neuro-symbolic integration, applying formal methods precisely where ensemble consensus is lowest, can improve the combined accuracy-plus-content-effect metric used by this task.

1 Introduction

Syllogistic reasoning is a fundamental form of deductive inference studied extensively in logic and cognitive science (Eisape et al., 2024; Bertolazzi et al., 2024). A key challenge in evaluating reasoning capabilities of language models is *content effects*: the tendency for real-world believability to interfere with purely logical judgment (Dasgupta et al., 2022). Mechanistic analyses have shown that language models develop reasoning circuits during pre-training, but these are susceptible to contamination from world knowledge (Kim et al., 2025), a phenomenon documented across tasks, model families, and domains (Wysocka et al., 2025; Ozeki et al., 2024). SemEval-2026 Task 11 (Valentino et al., 2026) formalizes this challenge by evaluating systems on syllogistic validity prediction using a combined metric that rewards both accuracy and low content effect.

In this paper, we describe FregeLogic, a hybrid neuro-symbolic system that exploits the complementary strengths of LLM ensembles and formal logic solvers. We construct a five-member ensemble from three open-weight model families (Llama 4 Maverick, Llama 4 Scout, Qwen3-32B) paired with varied prompting strategies, then use the Z3 SMT solver (De Moura and Bjørner, 2008) as a tiebreaker when the ensemble produces a narrow 3–2 vote split. The design is motivated by an empirical observation: close votes in the ensemble disproportionately coincide with content-biased errors, precisely the cases where a content-neutral formal verifier can add value (Bayless et al., 2025; Ranaldi et al., 2025). In nested 5-fold cross-validation on the dataset (N=960), this selective intervention reduces content effect by 16% (3.39 $\rightarrow$ 2.85) while improving accuracy by 0.9%, yielding a combined score of 41.88 compared to 39.12 for the pure ensemble.

2 Background

2.1 Task Description and Metrics

SemEval-2026 Task 11 (Valentino et al., 2026) evaluates syllogistic reasoning while disentangling logical structure from semantic content. We participate in Subtask 1: binary classification of syllogisms as valid or invalid. The dataset comprises 960 syllogisms annotated with validity labels and plausibility metadata (believable/unbelievable), balanced across four subgroups.

The task uses a combined score that jointly rewards high accuracy and low **content effect** (CE), which measures how much a system's predictions are influenced by believability rather than logical structure. The combined score applies a logarithmic penalty for content bias: Score $=$ Accuracy$/(1 + \ln(1 + \text{CE}))$. A content effect of zero indicates predictions uninfluenced by believability. Full metric definitions are in Valentino et al. (2026).

2.2 Related Work

Content effects in LLMs. Dasgupta et al. (2022) demonstrated that large language models exhibit human-like content effects across syllogism validity judgments and other reasoning tasks. Eisape et al. (2024) showed that even the largest models in the PaLM 2 family exhibit systematic biases including sensitivity to variable ordering, a structural bias related to, but distinct from, semantic content effects. Bertolazzi et al. (2024) found that pretrained LLM behavior can be explained by heuristics from cognitive science, and Ozeki et al. (2024) confirmed that LLMs exhibit human-like reasoning biases with primary limitations in the reasoning process itself.

Mechanistic interpretations. Kim et al. (2025) uncovered a three-stage reasoning circuit for syllogistic inference involving middle-term suppression and information propagation via mover heads. Critically, this circuit is susceptible to contamination from belief biases encoded in additional attention heads, providing mechanistic evidence for content effects and motivating our use of formal verification.

Neuro-symbolic approaches. Ranaldi et al. (2025) proposed quasi-symbolic chain-of-thought prompting to improve robustness on reasoning tasks. Bayless et al. (2025) presented a neurosymbolic framework for verifying logical correctness using SMT solvers, and Akinfaderin and Subramanian (2026) applied similar ideas to financial AI. Wysocka et al. (2025) showed that zero-shot LLMs achieve between 23% and 70% on biomedical syllogistic reasoning. Valentino et al. (2025) and Maraia et al. (2026) explored activation-level approaches to mitigate content effects. Our work differs from pure-LLM approaches by introducing structurally-grounded formal verification, from pure-formal approaches by handling paraphrase and extraction failures gracefully, from Bayless et al. (2025) by applying verification selectively only where ensemble consensus is low, and from activation-steering approaches (Valentino et al., 2025; Maraia et al., 2026) by requiring no access to model internals.

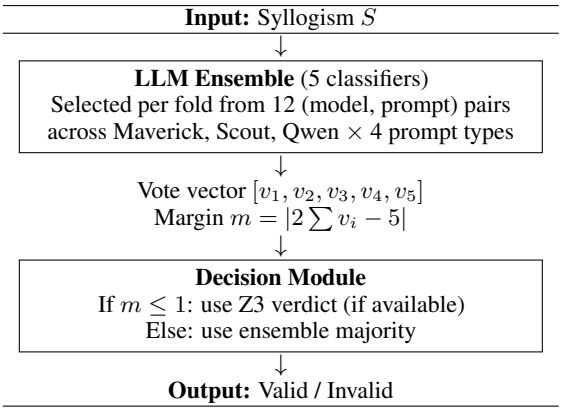

Figure 1: FregeLogic system architecture. The ensemble produces a vote vector; when the margin is low (3–2 split), the system defers to Z3 formal verification.

3 System Overview

Our system consists of three components: (1) an LLM ensemble that provides high-accuracy predictions through diverse model and prompt combinations, (2) a Z3-based formal verification pipeline that provides structurally-grounded logical judgments, and (3) a tiebreaker decision module that routes predictions to Z3 only when the ensemble produces a narrow vote margin (3–2 split). Figure 1 provides an overview of the architecture.

3.1 LLM Ensemble

The ensemble consists of five classifiers, each defined by a (model, prompt) pair. We use three open-weight models accessed via Amazon Bedrock: Llama 4 Maverick (17B active parameters, MoE architecture), Llama 4 Scout (17B), and Qwen3 (32B). These models were selected to maximize architectural diversity (MoE vs. dense, two distinct model families) within a practical parameter budget (17B–32B active parameters), while remaining accessible through a single inference API. They are paired with four prompting strategies to create diversity across both model family and reasoning elicitation method:

- **Zero-shot:** A direct validity question with no examples.
- **Few-shot:** Seven worked examples of valid and invalid syllogisms with explicit validity rules, balanced across plausibility conditions.

- **Few-shot CoT:** Worked examples with step-by-step reasoning traces demonstrating structural analysis, including a believable-invalid example that explicitly names the content bias.

- **Simple CoT:** A minimal prompt asking the model to identify logical structure before answering.

The five configurations are selected per fold based on combined score on a 200-sample inner subset (see Section 4). The selection naturally balances model and prompt diversity, following the principle that ensemble accuracy benefits from uncorrelated errors across members (Bertolazzi et al., 2024).

For each syllogism, all five classifiers produce a binary prediction $v_i \in \{0, 1\}$, where 1 indicates valid. The ensemble majority vote is $\hat{y} = \mathbb{1}[\sum_i v_i > 2.5]$.

All models are called with temperature 0.0 to maximize output reproducibility. Responses are parsed using a multi-stage regex pipeline that checks for explicit answer patterns (e.g., "ANSWER: true"), last-line heuristics, and fallback to the last occurrence of "true", "false", "valid", or "invalid" in the response. Parse failures default to invalid.

3.2 Z3 Formal Verification

The Z3 component (De Moura and Bjørner, 2008) provides structurally-grounded validity checking by encoding syllogisms in first-order logic (FOL) and testing satisfiability. We select Z3 over alternative provers for three reasons: (1) its native support for quantifiers and uninterpreted sorts allows direct encoding of syllogistic propositions without conversion to conjunctive normal form, unlike resolution-based provers such as Prover9 (Olausson et al., 2023); (2) its mature Python API enables tight integration with our LLM pipeline; and (3) its proven effectiveness in neuro-symbolic NLP, where it has been used alongside other solvers for faithful logical reasoning (Pan et al., 2023; Bayless et al., 2025). The pipeline consists of three stages.

Structure extraction. An LLM is prompted to extract the logical structure of the syllogism as a JSON object containing the three terms, two premises, and conclusion, each annotated with its proposition type (A: universal affirmative, E: universal negative, I: particular affirmative, O: particular negative) and its subject and predicate terms. We use Amazon Bedrock's structured output API with a JSON schema that enforces well-formed output, reducing extraction failures from approximately 22% (with free-form prompting) to near zero. Extraction is attempted with each of the three models in sequence (Maverick → Qwen → Scout, ordered by individual combined score), using the first successful parse. In practice, Maverick succeeds on 99.5–100% of cases across folds.

FOL encoding. Each proposition is encoded as a first-order formula over a declared sort `Thing` with unary predicates for each term. Based on diagnostic inspection of Felapton-type syllogisms in the dataset (all labelled valid), we adopt an Aristotelian interpretation with existential import, adding existence axioms ($\exists x : S(x)$) for subject terms in universal propositions:

- Type A ("All S are P"): $\forall x : S(x) \rightarrow P(x) \wedge \exists x : S(x)$
- Type E ("No S are P"): $\forall x : S(x) \rightarrow \neg P(x) \wedge \exists x : S(x)$
- Type I ("Some S are P"): $\exists x : S(x) \wedge P(x)$
- Type O ("Some S are not P"): $\exists x : S(x) \wedge \neg P(x)$

Satisfiability check. Validity is determined by a two-step procedure. First, we verify premise consistency: if $P_1 \wedge P_2$ is UNSAT (e.g., a Type I premise paired with a contradicting Type E premise over the same terms), the premises are mutually inconsistent and Z3 returns $\perp$, deferring to the ensemble. This guards against *ex contradictione quodlibet* false positives, where any conclusion would be trivially entailed by inconsistent premises. Otherwise, validity is checked in the standard way:

$$\text{valid}(S) \iff (P_1 \wedge P_2 \text{ is SAT}) \wedge (P_1 \wedge P_2 \wedge \neg C \text{ is UNSAT}) \quad (1)$$

If the second check returns UNSAT, the syllogism is valid; if SAT, it is invalid; if UNKNOWN (timeout at 5000ms), the result is undefined ($\perp$). When structure extraction fails or produces malformed output, Z3 also returns $\perp$.

The Z3 *solver* is content-neutral by construction, since the formal encoding strips away all semantic content. However, the end-to-end pipeline's content independence is contingent on the extraction step. Our extraction failure disaggregation (Section 5) confirms that failure rates are uniform across

plausibility subgroups. Pipeline accuracy depends entirely on correct structure extraction, which can fail when the LLM misidentifies proposition types or term boundaries.

3.3 Tiebreaker Decision Fusion

The decision module combines ensemble and Z3 predictions based on vote consensus, measured by the vote margin:

$$m(S) = |2 \cdot \textstyle\sum_i v_i - n| \tag{2}$$

where $n = 5$ is the number of classifiers. For $n = 5$, the margin is 5 (unanimous), 3 (4–1 split), or 1 (3–2 split).

The final prediction is:

$$\text{pred}(S) = \begin{cases} z(S) & m(S) \leq 1 \wedge z(S) \neq \perp \\ \hat{y}(S) & \text{otherwise} \end{cases} \tag{3}$$

where $z(S)$ is the Z3 verdict and $\hat{y}(S)$ is the ensemble majority vote. The threshold $\tau = 1$ means Z3 is consulted only on 3–2 splits, which we hypothesize correspond to cases where content bias causes disagreement among the LLMs.

4 Experimental Setup

The dataset provided by the task organizers (Valentino et al., 2026) contains N=960 syllogisms annotated with validity labels and plausibility metadata, balanced across four subgroups: valid-believable (240), valid-unbelievable (240), invalid-believable (234), and invalid-unbelievable (246). No separate development set was provided.

We adopt a nested 5-fold cross-validation protocol to separate model selection from evaluation. In each outer fold, 768 syllogisms are used for calibration and 192 for evaluation. Within each calibration set, a 200-sample inner subset is randomly drawn for Phase 1 model selection, where all 12 combinations of 3 models and 4 prompts are scored. Two additional OpenAI models (GPT-OSS-120B and GPT-OSS-20B) were evaluated in preliminary experiments but excluded due to low combined scores (14.7–21.0; see Appendix C). The top five configurations by combined score are selected per fold to form the ensemble. All LLM calls use Amazon Bedrock with temperature 0.0. The Z3 solver uses the Python API with a 5000ms timeout. Full prompts are in Appendix A.

Strategy	Acc	CE	Score
Ensemble (pure)	93.4±1.5	3.39±1.30	39.12±5.37
+ Z3 Tiebreaker	**94.3±0.9**	**2.85±1.23**	**41.88±5.97**
+ Z3 Weighted	93.4±1.5	3.39±1.30	39.12±5.37
+ Z3 Veto	74.1±2.1	26.70±2.97	17.18±0.84
Confidence + Z3	91.7±1.1	6.15±2.00	31.77±3.82
Top 3 + Z3	92.2±2.0	5.08±2.25	34.11±4.36
Z3 Only	74.7±2.0	26.28±2.88	17.39±0.87

Table 1: Comparison of fusion strategies (mean ± std across 5 outer folds). The tiebreaker strategy achieves the best combined score. Bold indicates the best result.

Existential import. To determine the correct FOL encoding, we inspected the dataset for Darapti-type (AAI-3) and Felapton-type (EAO-3) syllogisms, whose validity depends on whether existential import is assumed. No Darapti candidates were found; seven Felapton candidates were found, all labelled valid. This confirms an Aristotelian interpretation, and we add existence axioms accordingly (Section 3.2).

5 Results and Analysis

5.1 Individual Model Performance

Table 5 (Appendix D) presents results for all 12 model-prompt combinations from a representative Phase 1 inner fold. Accuracy varies modestly across configurations (79.5–89.0%), but content effect varies substantially (3.49–13.71), driving large differences in combined score. The best prompt strategy differs by model: Maverick favours zero-shot and few-shot, while Qwen benefits from chain-of-thought. Notably, Simple CoT achieves the lowest average content effect (5.24) despite being the most under-specified prompt, suggesting that minimal instructions may allow models to engage internal structural reasoning rather than anchoring to example content. Chain-of-thought does not uniformly reduce content effects; for Maverick, few-shot CoT increases CE from 3.83 to 10.11 compared to plain few-shot.

5.2 Ensemble and Hybrid Strategies

Table 1 compares six fusion strategies aggregated across the five outer folds. The pure ensemble achieves 93.4% accuracy but a content effect of 3.39. The Z3-only baseline achieves low accuracy (74.7%) with a high content effect (26.28), driven by systematic over-prediction of invalidity.

The tiebreaker strategy achieves the best combined score (41.88) by reducing the content effect from 3.39 to 2.85, a 16% reduction, while also im-

Strategy	VB	VU	IB	IU
Ensemble (pure)	95.9	96.0	90.2	91.9
+ Z3 Tiebreaker	95.6	93.8	94.5	93.5
Z3 Only	50.5	53.9	98.0	97.2

Table 2: Subgroup accuracy (%) for key strategies, averaged across folds. VB = Valid-Believable, VU = Valid-Unbelievable, IB = Invalid-Believable, IU = Invalid-Unbelievable.

proving accuracy by 0.9 percentage points. This is consistent with our hypothesis that ensemble disagreements correspond to content-biased cases that Z3 can help correct.

Table 2 provides subgroup accuracy for the key strategies. The tiebreaker achieves more balanced accuracy across all four subgroups than either the pure ensemble or Z3 alone. The pure ensemble shows a gap between valid subgroups (95.9–96.0%) and invalid subgroups (90.2–91.9%); the tiebreaker narrows this gap (93.5–95.6%), with the largest gains on invalid-believable cases (90.2% $\rightarrow$ 94.5%) where content bias is strongest.

The remaining strategies perform worse: Z3 veto (score 17.18) overrides correct high-consensus predictions, the weighted strategy never flips a majority, and the confidence-based strategy (91.7%) trusts Z3 on both 3–2 and 4–1 splits (effectively lowering the margin threshold to τ=3), overriding the ensemble on 37–52 cases per fold compared to 13–19 for the tiebreaker. Because Z3 accuracy on valid syllogisms is low (48.6% on 3–2 split cases; see Section 5.4), extending Z3 authority to higher-consensus cases where the ensemble is typically correct degrades overall performance.

5.3 Analysis of Tiebreaker Behavior

Table 3 summarizes the tiebreaker mechanism's behavior aggregated across all five folds. The 3–2 vote split occurs in 7.9% of cases (76 of 960). Z3 produces a usable verdict on all 76 cases, enabled by structured-output extraction which reduced failures to near zero. Of the 30 cases where Z3 overrides the ensemble majority, 19 are correct flips and 11 are wrong flips, yielding a net improvement of 8 correct predictions.

When all five LLMs agree, the logical signal is typically strong enough to overcome content bias. The 3–2 splits indicate cases where some models are swayed by content while others are not; Z3 resolves these independently of semantic content. Coalition analysis reveals that Scout appears in the minority coalition on 53.9% of 3–2 splits (41 of 76), above the 40% expected by chance, suggesting greater susceptibility to content bias. Conversely, Maverick+FS appears in the minority least often (28.9%, 22 of 76), well below the 40% chance baseline, indicating the strongest resistance to content effects among the five classifiers.

Metric	Count	%
Total evaluation instances	960	100
5–0 or 4–1 splits	884	92.1
3–2 splits (tiebreaker triggered)	76	7.9
Z3 available on splits	76	100.0
Degenerate premises	1	0.1
Z3 override decisions	30	
Correct flips	19	
Wrong flips	11	

Table 3: Tiebreaker behavior aggregated across all five outer folds.

5.4 Z3 Invalidity Bias

The Z3-only baseline exhibits a pronounced invalidity bias: it achieves 97.6% accuracy on invalid syllogisms (IB: 98.0%, IU: 97.2%) but only 52.2% on valid ones (VB: 50.5%, VU: 53.9%). Because the Z3 solver itself is deterministically correct given well-formed input, this asymmetry originates in the extraction step. When the LLM misidentifies a proposition type or term boundary in a valid syllogism, the resulting FOL encoding almost always breaks the entailment chain, causing Z3 to correctly report unsatisfiability of the (now-corrupted) logical structure. Extraction errors on invalid syllogisms are less consequential: the syllogism is already non-entailing, so a different structural error is unlikely to accidentally produce a valid derivation. This directional asymmetry explains both the Z3-only subgroup collapse on valid cases and why the tiebreaker restricts Z3 authority to low-consensus cases, where the benefit of correcting content-biased ensemble errors outweighs the risk of extraction-induced false invalidity verdicts.

5.5 Error Analysis

Three main error sources remain: (1) **Z3 extraction errors**, where incorrect structure extraction (e.g., misidentified proposition types) produces wrong Z3 verdicts, accounting for 11 wrong flips across all folds; (2) **unanimous ensemble errors**, where all five LLMs agree incorrectly on syllogisms with strong content bias, preventing tiebreaker intervention; and (3) **degenerate premises**, where mutually inconsistent premises trigger the consistency check and Z3 defers to the ensemble (1 case across all

(a) Proposition type error *AAI-1, unbelievable*

"There are no humans that are not made of glass. Every single thing made of glass is indestructible. ∴ some indestructible things are humans."

True: Valid **Z3:** Invalid **Cause:** Double negative "no . . . not" → type E instead of A

(b) Term boundary error *Darii IAI-3, unbelievable*

"Some fish are penguins. All penguins belong to the class of birds. ∴ some birds are fish."

True: Valid **Z3:** Invalid **Cause:** "belong to the class of" → splits "class of birds" as compound term

(c) Non-standard quantifier *Felapton EAO-1, believable*

"Anything that is a rose is a flower. Under no circumstances is a flower a tree. ∴ some trees are not roses."

True: Valid **Z3:** Invalid **Cause:** "Under no circumstances" → may invert subject-predicate or miss E-type

Figure 2: Three representative wrong flips. All 11 go in the same direction: Z3 falsely rejects valid syllogisms due to extraction errors, not encoding errors.

folds).

Extraction failure disaggregation confirms that the structured-output pipeline is content-neutral: failure rates are 0.0% for VB, IB, and IU subgroups, and 0.4% for VU (1 of 241 cases), consistent with uniform distribution.

All 11 wrong flips share the same direction: Z3 falsely rejects a valid syllogism. Eight involve unbelievable content, where absurd premises may compound extraction difficulty. Figure 2 presents three representative cases. In each, the Z3 solver produces a correct verdict given the (incorrectly) extracted structure, confirming that the bottleneck is extraction fidelity, not logical encoding.

Additional analysis of prompt strategies is in Appendix E.

6 Conclusion

We presented FregeLogic, a hybrid neuro-symbolic system that combines an LLM ensemble with Z3 formal verification via a tiebreaker mechanism for syllogistic validity prediction. Our results support the hypothesis that ensemble disagreement signals content-biased predictions, which formal logic can help correct. In nested 5-fold cross-validation on the dataset (N=960), the tiebreaker achieves a combined score of 41.88, a 2.76-point improvement over the pure ensemble driven by a 16% reduction in content effect and a 0.9% accuracy gain. Adopting structured-output API calls for extraction reduced Z3 failure rates from approximately 22% to near zero, and a two-step satisfiability check with Aristotelian existential import was validated against task annotations. Future work includes investigating whether the tiebreaker mechanism generalizes to other reasoning tasks and exploring adaptive thresholds for the vote margin cutoff.

Limitations

While our approach involves no parameter updates, model and prompt selection, fusion strategy selection, and the tiebreaker threshold (τ=1) were all tuned via nested cross-validation on the provided dataset; no independent test set was used. This selection procedure adds non-trivial setup complexity: each fold requires scoring all 12 (model, prompt) combinations on a 200-sample inner subset before the top-five ensemble is fixed, and the optimal configuration differs across folds (Appendix D), so the approach does not reduce to a single off-the-shelf recipe. The Z3 pipeline depends on LLM-based structure extraction, which, despite structured-output enforcement, can still produce semantically incorrect extractions (e.g., misidentified proposition types), accounting for the 11 wrong flips observed. Our system requires six LLM API calls per syllogism (five ensemble plus one extraction) and one Z3 solve, totalling approximately 4,300 tokens and 12.1 seconds of sequential latency per instance (2.8 seconds with parallel ensemble calls); Z3 solving adds under 10ms. Full cost profiling over 18,722 API calls is reported in Appendix G. We did not compare against larger single dense models (e.g., 70B+); whether architectural diversity at smaller scale outperforms a single larger model on this task remains an open question for future work.

Acknowledgments

We thank the SemEval-2026 Task 11 organizers for designing a task that foregrounds the important challenge of content effects in reasoning evaluation.

References

Adewale Akinfaderin and Shreyas Subramanian. 2026. VERAFI: Verified agentic financial intelligence through neurosymbolic policy generation. In *AAAI 2026 Workshop on Agentic AI in Financial Services*.

Sam Bayless, Stefano Buliani, Darion Cassel, Byron Cook, Duncan Clough, R'emi Delmas, Nafi Diallo,

Ferhat Erata, Nick Feng, Dimitra Giannakopoulou, and 1 others. 2025. A neurosymbolic approach to natural language formalization and verification. *arXiv preprint arXiv:2511.09008*.

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 Processing*, pages 13882–13905.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: An efficient SMT solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pages 337–340. Springer.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van 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*.

Geonhee Kim, Marco Valentino, and Andr'e Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095.

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

Theo X. Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. 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.

Kazuki Ozeki, Risako Ando, Takuro Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang 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.

Leonardo Ranaldi, Marco Valentino, and Andr'e Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and Andr'e Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *arXiv preprint arXiv:2505.12189*.

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)*.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andr'e Freitas. 2025. SylloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 7235–7258.

A Prompt Templates

We provide the four prompt templates used in our system. In each template, `{syllogism}` is replaced with the input syllogism text.

A.1 Zero-Shot Prompt

```
Determine if this syllogism is VALID.

VALID means: IF the premises were true, the conclusion MUST be true. Ignore whether premises are actually true in the real world.

Syllogism: {syllogism}

Answer with exactly one word: true or false
```

A.2 Few-Shot Prompt

```
Determine if this syllogism is VALID (conclusion necessarily follows from premises).

VALIDITY RULES:
- “All A are B” + “All B are C” -> “All A are C” (valid)
- “No A are B” + “All C are A” -> “No C are B” (valid)
- “All A are B” + “Some C are A” -> “Some C are B” (valid)
- “All A are B” + “All C are B” -> “All A are C” (invalid, undistributed middle)
- “Some A are B” does NOT guarantee “All A are B”

EXAMPLES:
“All dogs are mammals. All mammals are animals. Therefore, all dogs are animals.” -> true
“All birds are dinosaurs. All sparrows are birds. Therefore, all sparrows are
```

dinosaurs." -> true
"No fish are mammals. All sharks are fish. Therefore, no sharks are mammals." -> true
"All reptiles are cold-blooded. Some lizards are reptiles. Therefore, some lizards are cold-blooded." -> true
"All lawyers are professionals. All doctors are professionals. Therefore, all lawyers are doctors." -> false
"Some politicians are corrupt. All senators are politicians. Therefore, some senators are corrupt." -> false
"All rocks are edible. Some clouds are rocks. Therefore, all clouds are edible." -> false

Syllogism: {syllogism}

Answer with exactly one word: true or false

A.3 Few-Shot Chain-of-Thought Prompt

Analyze this syllogism's logical VALIDITY.

IMPORTANT: VALID = conclusion MUST follow IF premises are assumed true. Ignore real-world facts.

RULES:
- "All A are B" + "All B are C" -> "All A are C" (valid chain)
- "No A are B" + "All C are A" -> "No C are B" (valid exclusion)
- "All A are B" + "Some C are A" -> "Some C are B" (Darii)
- "All A are B" + "All C are B" -> "All A are C" (invalid, undistributed middle)
- "Some A are B" means ONLY SOME, not all

WORKED EXAMPLES:

Example 1: "All cats are mammals. All mammals are animals. Therefore, all cats are animals."
- Structure: cats $\subseteq$ mammals $\subseteq$ animals
- Chain is complete. ANSWER: true

Example 2: "All unicorns fly. All pegasi are unicorns. Therefore, all pegasi fly."
- Premises are fantasy but structure is: pegasi $\subseteq$ unicorns $\subseteq$ fly
- Valid chain regardless of real-world truth. ANSWER: true

Example 3: "All athletes are healthy. All healthy people exercise. Therefore, all athletes exercise."
- Chain: athletes -> healthy -> exercise
- Chain is complete. ANSWER: true

Example 4: "All doctors are professionals. All lawyers are professionals. Therefore, all doctors are lawyers."
- "Professionals" appears as PREDICATE in both premises.
- Middle term is undistributed: we only know both are subsets of professionals, not that they overlap.
- Despite the believable surface, the structure is invalid. ANSWER: false

Example 5: "All cats are pets. All dogs are pets. Therefore, all cats are dogs."
- Both subsets of pets, but could be separate
- Undistributed middle. ANSWER: false

Example 6: "Some birds can fly. All penguins are birds. Therefore, some penguins can fly."
- "Some birds" doesn't tell us WHICH birds
- Cannot guarantee any penguin is in the flying subset. ANSWER: false

Syllogism: {syllogism}

Think through the structure briefly, then write your final answer as: ANSWER: true or ANSWER: false

A.4 Simple Chain-of-Thought Prompt

Is this syllogism logically VALID? (If premises were true, must conclusion be true?)

Syllogism: {syllogism}

First, identify the logical structure. Then determine if the conclusion necessarily follows.

End your response with exactly: ANSWER: true or ANSWER: false

B Z3 Structure Extraction Prompt

The following prompt is used to extract syllogistic structure for Z3 encoding. When supported, extraction uses Bedrock's structured output API with a JSON schema enforcing the expected format; otherwise, the model's free-form response is parsed.

Extract the logical structure of this syllogism.

SYLLOGISM: {syllogism}

Proposition types:
- A: "All S are P" / "Every S is P"
- E: "No S are P"
- I: "Some S are P" / "At least one S is P"
- O: "Some S are not P"

The CONCLUSION follows "therefore/hence/thus/consequently/so".

Output ONLY this JSON (replace t1/t2/t3 with the exact term WORDS from the syllogism text):
{"terms": ["t1", "t2", "t3"],
"premise1": {"type": "A/E/I/O",
"subject": "term <- exact word(s) from text",
"predicate": "term <- exact word(s) from text"},
"premise2": {"type": "A/E/I/O",
"subject": "term <- exact word(s) from text",
"predicate": "term <- exact word(s) from text"},
"conclusion": {"type": "A/E/I/O",

Model	Prompt	Acc	CE	Score
GPT-120B	Zero-Shot	80.0	15.71	20.96
GPT-120B	Few-Shot	78.0	18.64	19.61
GPT-120B	FS-CoT	76.0	20.32	18.72
GPT-120B	Simple CoT	75.5	19.88	18.69
GPT-20B	Zero-Shot	79.0	13.67	21.43
GPT-20B	Few-Shot	78.5	14.71	20.91
GPT-20B	FS-CoT	73.0	18.92	18.29
GPT-20B	Simple CoT	65.5	31.14	14.65

Table 4: Results for excluded models on a 200-sample evaluation set. These models exhibited higher content effects on average than the selected models, particularly for chain-of-thought prompts, resulting in lower combined scores across all four prompt types.

Model	Prompt	Acc	CE	Score
Maverick	Zero-Shot	86.0	3.49	34.38
Maverick	Few-Shot	88.0	3.83	34.17
Maverick	FS-CoT	88.5	10.11	25.97
Maverick	Simple CoT	84.0	5.86	28.71
Scout	Zero-Shot	81.0	13.71	21.96
Scout	Few-Shot	89.0	6.92	29.00
Scout	FS-CoT	88.5	11.21	25.27
Scout	Simple CoT	85.0	4.79	30.84
Qwen	Zero-Shot	80.5	8.86	24.48
Qwen	Few-Shot	79.5	11.37	22.62
Qwen	FS-CoT	89.0	5.20	31.52
Qwen	Simple CoT	88.0	5.07	31.39

Table 5: Individual model-prompt results from a representative Phase 1 inner fold (200 samples). Acc = Accuracy (%), CE = Content Effect, Score = Combined Score. FS-CoT = Few-Shot Chain-of-Thought.

```
"subject": "term <- exact word(s) from
text",
"predicate": "term <- exact word(s) from
text"}}
```

C Excluded Models

Table 4 reports results for GPT-OSS-120B and GPT-OSS-20B, which were evaluated but excluded from the final ensemble due to low combined scores driven by high content effects.

D Individual Model Results

Table 5 presents results for all 12 model-prompt configurations from a representative Phase 1 inner fold (200 samples).

E Prompt Strategy Analysis

Table 6 shows average metrics by prompt type across the three selected models on a representative 200-sample inner fold. Simple CoT achieves the best average combined score (30.31) due to its low content effect (5.24), despite not having the highest accuracy. Few-shot CoT increases content effect for Maverick and Scout while decreasing it for Qwen, indicating an interaction between model architecture and reasoning elicitation. This suggests that minimal reasoning prompts may be preferable to elaborate demonstrations when the goal is content-independent reasoning.

Prompt Type	Avg Acc	Avg CE	Avg Score
Zero-Shot	82.50	8.69	26.94
Few-Shot	85.50	7.37	28.60
Few-Shot CoT	88.67	8.84	27.59
Simple CoT	85.67	5.24	30.31

Table 6: Average metrics by prompt type across three models on a representative 200-sample inner fold.

F Strategy Trade-offs

Figure 3 shows subgroup accuracy for the three key strategies. The tiebreaker (solid green) achieves the most symmetric profile, while the pure ensemble (dashed blue) favours valid subgroups and Z3 Only (dotted red) favours invalid subgroups.

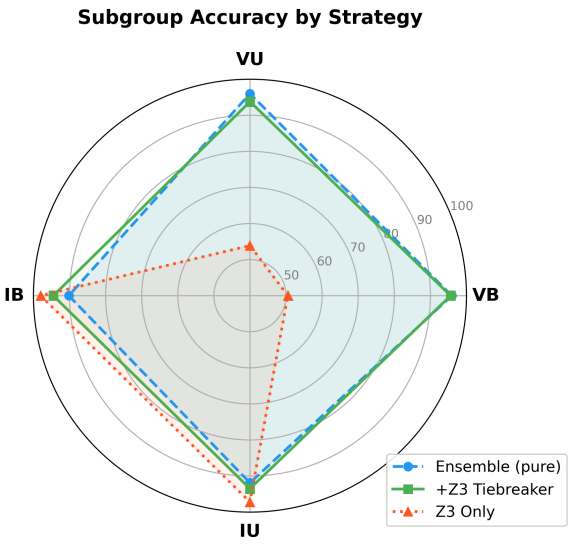

Figure 3: Subgroup accuracy by strategy. The tiebreaker achieves the most balanced profile across all four subgroups.

G Computational Cost

Table 7 reports per-component cost profiling from the full nested cross-validation run (18,722 API calls across model selection and evaluation phases). Qwen3-32B is the fastest model per call (median 1.1s), while Maverick and Scout average 2.5–2.8s at the median. Z3 solving is negligible: 960 calls complete in 8.5 seconds total (median 7ms). At inference time, each syllogism requires five ensemble calls, one extraction call, and one Z3 solve,

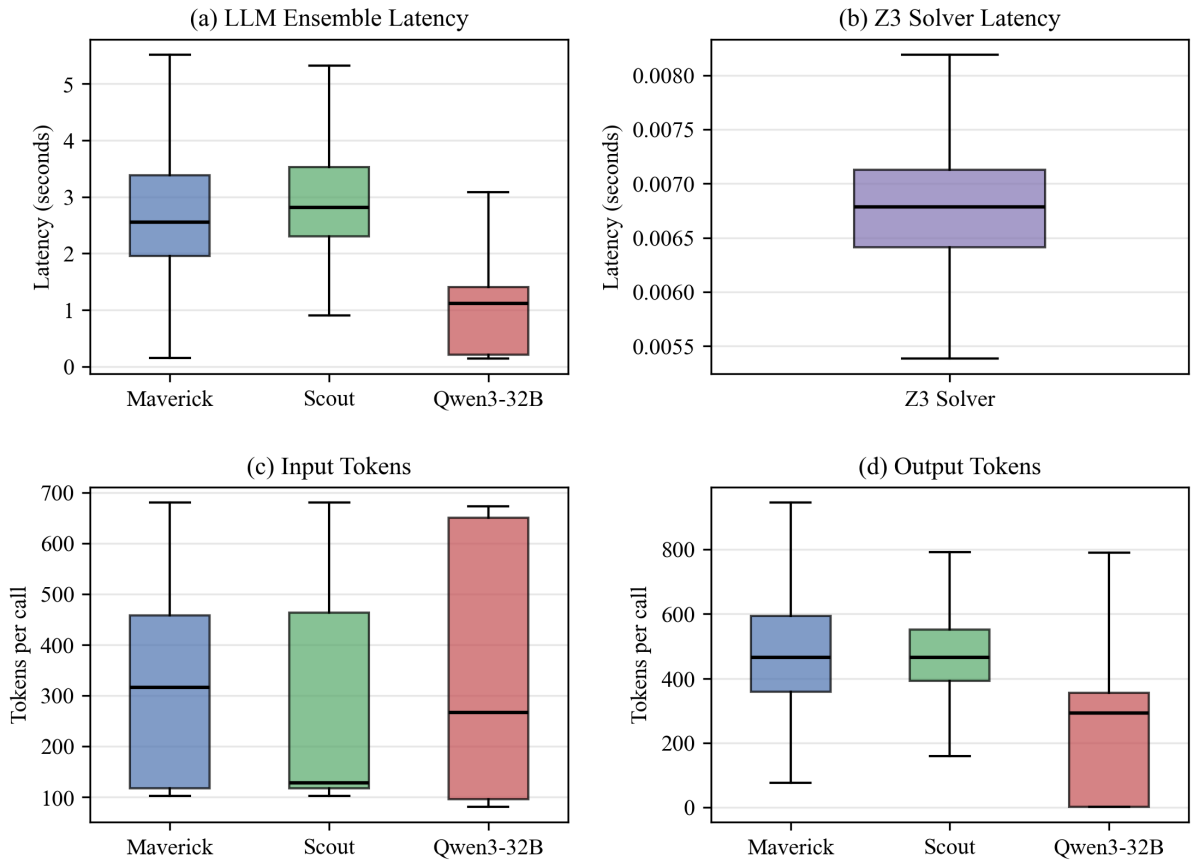

Figure 4: Per-call cost distributions across the full cross-validation run (18,722 calls). (a) LLM ensemble latency, (b) Z3 solver latency (note the millisecond scale), (c) input tokens per call, (d) output tokens per call. Outliers suppressed for readability. The computational bottleneck is LLM inference, not formal verification.

Component	n	**p50 (s)**	**p95 (s)**	**Tokens**
Maverick (ens.)	7,264	2.55	5.25	5.84M
Scout (ens.)	4,576	2.82	5.26	3.69M
Qwen3-32B (ens.)	5,920	1.11	1.86	3.35M
Qwen3-32B (extr.)	2	0.49	0.49	1.4K
Z3 Solver	960	0.007	0.009	—
Total	**18,722**			**12.87M**

Table 7: Per-component cost profiling across the full nested 5-fold cross-validation run. n = number of calls, p50/p95 = median and 95th-percentile latency, Tokens = total input + output tokens. Extraction calls are rare because Maverick succeeds on $>$99.5% of cases.

consuming approximately 2,200 input tokens and 2,100 output tokens.

Figure 4 shows per-call distributions across all four dimensions. LLM ensemble calls range from 1 to 5 seconds, with Qwen3-32B consistently faster than the Llama 4 models but consuming more input tokens due to longer prompt encodings. Maverick and Scout produce more output tokens on average (488 and 480 vs. 228 for Qwen), reflecting more verbose chain-of-thought responses. Z3 solve times are three orders of magnitude smaller than any LLM call, confirming that the formal verification step adds negligible overhead.