

Neuro-Symbolic Reconstruction of Philosophical and Formal Arguments: Automated Propositional Mapping and Verdict Verification via LLMs and SMT Solvers

Arian Esmailinasab¹ | Amir A. Ghavifekr^{✉ 2} 

1. Department of Control Engineering, University of Tabriz, Tabriz, Iran. E-mail: arian.esmailinasab@ieee.org

2. Corresponding Author, Assistant Professor of Control Engineering Department, University of Tabriz, Tabriz, Iran. E-mail: aa.ghavifekr@tabrizu.ac.ir

Article Info

Article type:

Research Article

Article history:

Received 01 October 2026

Received in revised form
20 October 2026

Accepted 24 July 2026

Published online 25
October 2026

Keywords:

Argument Reconstruction,
Enthymemes, First-order
Logic, Large Language
Models, Neuro-Symbolic
Systems, Z3 SMT solver

ABSTRACT

“To test whether an end-to-end neuro-symbolic pipeline (deterministic segmentation, LLM formalisation into SMT-LIB2, automated symbolic repair, and three-way Z3 SMT verification with countermodel generation) can reconstruct classical philosophical arguments, and to quantify where scaling changes the character of formalisation failures.” Five-stage pipeline evaluated on 31 curated philosophical and explicit-logic arguments plus the full 204-item FOLIO validation split under a three-way protocol, replicated across three independently pinned backends (Qwen3-4B-Instruct; Gemini 3.8 Flash; Nemotron-3-Ultra-550B). An independent audit script re-derives every reported number from the archived raw records. Explicit logic vs. philosophical prose: 68.8%/27.3% (Qwen) and 93.8%/63.6% (Gemini); replicated effect sizes of -41.5 and -30.2 percentage points (Fisher $p = 0.054/0.125$, not significant for the larger model). Enthymemes: 0.0% vs 50.0%; a warrant-eliciting prompt repairs the barrier only for the larger model (7/11 to 10/11) at a cost on explicit items (8/8 to 6/8). FOLIO: a refutation asymmetry replicates in all three families and decided-item refutation never exceeds 58.1% (36/62); the direct no-solver baseline is 66.7% for the small model. LLMs transcribe fluently but reconstruct poorly; the deficit on classical prose is necessary rather than technical. The pipeline's value is inspectable judgment — solver-checked verdicts, per-stage diagnosis, and countermodels as automated thought experiments.

Cite this article: Esmailinasab, A. & Ghavifekr, A. A. (2026). Neuro-Symbolic Reconstruction of Philosophical and Formal Arguments: Automated Propositional Mapping and Verdict Verification via LLMs and SMT Solvers. *Journal of Philosophical Investigations*, 20(56), 845-856. <https://doi.org/10.22034/jpiut.2026.22498>



© The Author(s).

Publisher: University of Tabriz.

Introduction

Recent breakthroughs in large language models (LLMs) have motivated substantial interest in their capacity for formal logical deduction and automated reasoning. Neuro-symbolic frameworks, which couple neural generative flexibility with the deterministic soundness of Satisfiability Modulo Theories (SMT) solvers and automated theorem provers, have achieved impressive scores on curated benchmarks such as FOLIO (Han et al., 2024) and ProofWriter (Tafjord et al., 2021), and neuro-symbolic systems such as Logic-LM (Pan et al., 2023) and SatLM (Ye et al., 2023) route the model's output through a symbolic checker. In these environments, neural models translate structured natural language statements into first-order logic (FOL), which is subsequently resolved by an SMT solver.

Yet these benchmarks clean up the problem space artificially. Inputs arrive pre-segmented into neat atomic propositions; variables are unambiguously named, and inferential steps proceed without unstated logical leaps. Authentic philosophical texts do not obey these textbook conventions. As Walton et al. (2008) observe, human argumentation is fundamentally enthymematic: writers routinely omit metaphysical bridge principles, dialectical warrants, and shared background axioms.

Feed an unsupplemented enthymeme directly into an SMT solver, and the output is mathematically predictable: the solver returns invalid. What this verdict signifies remains systematically unclear. It could indicate a genuine logical fallacy by the philosopher; it could signify an incomplete translation by the neural model; or it may simply reveal the expressibility boundary of first-order solvers. To untangle these possibilities, we deploy an inspectable neuro-symbolic pipeline that formalizes classical arguments into SMT-LIB2, runs automated syntactic repair, and uses Z3 to produce both validity verdicts and explicit countermodels. The primary contributions of this paper are:

- 1) **An End-to-End Neuro-Symbolic Pipeline:** We design a modular pipeline featuring deterministic discourse segmentation, constrained LLM formalisation, automated symbolic repair, and Z3 SMT verification.
- 2) **Three-Way SMT Verification:** We establish an automated three-way decision procedure (True, False, Uncertain) that effectively differentiates entailment from refutation.
- 3) **Multi-Scale Empirical Evaluation:** We benchmark a 4B parameter open-weights model (Qwen3-4B) against a larger API model (Gemini 3.8 Flash) across 31 curated philosophical and explicit logic arguments, plus the full 204-item FOLIO validation split under three independently pinned backends from different model families.
- 4) **it will be demonstrated that the larger API model reaches a 100% parse rate on the curated corpus, but dialectical enthymemes remain largely unrecovered in both models (50.0% vs 0.0%), and a warrant-eliciting prompt can repair it just for larger API model.**

1. Neuro-Symbolic Pipeline Architecture

The proposed structure consists of five sequential phases, where the generative variance is confined strictly to propositional mapping.

1-1 Deterministic Discourse Segmentation

Regarding of using stochastic LLMs for text segmentation, a deterministic lexical partition has been utilized in stage 1. An ordered lexicon of causal discourse connectives ('therefore', 'hence', 'it follows that', and Persian markers) identifies premise propositions and the conclusion C, preventing compounding segmentation errors.

1.2 Neural Formalisation to SMT-LIB2

The LLM translates the natural-language propositions into standard SMT-LIB2. The prompt enforces explicit sort declarations, predicate signatures (declare-fun P (Int) Bool), individual constants (declare-const a Int), and first-order quantifiers (forall, exists).

1-3 Automated Symbolic and Syntactic Repair

Language models usually neglect variable declarations or produce trailing parenthesis mismatches on deeply nested quantifiers. Our symbolic repair layer:

- 1) Scans assertions for undeclared predicate and constant symbols;
- 2) Excludes quantifier-bound variables (e.g., x in (forall ((x Int)) (P x))) to avoid false declarations;
- 3) Synthesizes and prepends missing signatures ((declare-fun Q (Int) Bool));
- 4) Balances closing parentheses on assertion lines, preventing solver syntax exceptions.

1-4 SMT Solving and Countermodel Extraction

Stage 4 executes the standalone Z3 SMT solver binary, version 5.3.1 (64-bit), with a 60-second timeout per query (de Moura & Bjørner, 2008). Entailment of C from the premises is tested via refutation by checking satisfiability of the premise set together with the negated conclusion. If Z3 returns 'unsat', the argument is formally valid. If Z3 returns 'sat', the argument is invalid, and Z3 synthesizes a concrete countermodel an assignment of constants representing a possible world wherein all premises hold true while the conclusion is false.

1-5 Three-Way Verdict Generation

To capture nuances masked by binary validity, Stage 5 symmetrically queries Z3:

- Entailed (True): the premises conjoined with the negated conclusion is UNSAT.
- Refuted (False): the premises with the negated conclusion is SAT, and the premises with the conclusion is UNSAT.
- Undetermined (Uncertain): both conjoined sets are SAT.

2. Experimental Evaluation and Results

2-1 Curated Benchmark Corpus

The architecture has been evaluated on 31 curated arguments comprising:

- Explicit Logic (n=16): Fully specified first-order syllogisms modeled after FOLIO.
- Classical Philosophical Prose (n=15): Authentic historical arguments from Avicenna, al-Ghazali, Mulla Sadrā, al-Kindi, Aquinas, Plato, Kant, Ibn Rushd, Descartes, Hume, and Wittgenstein.

2-2 Model Comparison and Scaling Effects

We compared Qwen3-4B-Instruct (local 4B open weights) against Gemini 3.8 Flash (larger API model) at temperature 0. Table 1 summarizes headline performance across attempted items (95% bootstrap CIs).

Table 1. Overall Benchmark Performance (N=31)

Model	Scale	Parse Rate	Repair Rate	Validity Acc. [95% CI]
Qwen3-4B-Instruct	4B (Local)	93.5%	9.7%	51.9% [33.3%, 70.4%]
Gemini 3.8 Flash	Larger (API)	100.0%	6.5%	81.5% [66.7%, 96.3%]

The larger model raises validity accuracy by 29.6 percentage points (51.9% → 81.5%) and achieves a 100.0% parse rate on this corpus (with n = 31 the exact upper bound still admits ~10% parse failures, and on FOLIO prose the same model makes 6 solver or parse errors). Exact binomial (Clopper–Pearson) intervals, recomputed independently by the audit script, give [31.9%, 71.3%] and [61.9%, 93.7%], corroborating the bootstrap bands. A majority baseline must also be reported: 21 of the 27 graded items are gold-valid, so an always-Valid strategy scores 77.8% above the small model's 51.9%, and the benchmark information therefore lives in the paired per-item structure and the group gaps, not in the headline rates.

2-3 The Domain Disparity: Explicit Logic vs. Philosophical Prose

Table 2 and Fig. 2 reveal a striking domain gap: on explicit logic, Gemini 3.8 Flash achieves 93.8% and Qwen3-4B achieves 68.8%. On philosophical prose, accuracy collapses to 63.6% and 27.3% respectively. Of the 15 philosophical arguments, 4 are contested (not first-order expressible, Section IV.C) and are excluded from scoring, leaving n=11 graded; 4 of those 11 carry a suppressed warrant.

Table 2. Performance Breakdown Across Argument Domains

Model	Explicit (n=16)	Philosophical (n=11)	Domain Gap	Fisher exact (p)
Qwen3-4B-Instruct	68.8%	27.3%	-41.5 pp	p = 0.054 (n.s.)
Gemini 3.8 Flash	93.8%	63.6%	-30.2 pp	p = 0.125 (n.s.)

Because both models answer the identical 27 graded items, the scale comparison is stronger than its cell counts suggest: the two disagree on 10 items, 9–1 in the larger model's favour (exact two-sided sign test, p = 0.021; Fig. 1). Consistent with the exact 2×2 tests in Table 2 — the appropriate primary statistic at these cell sizes neither gap reaches significance on its own; we

therefore state the finding as a replicated effect size (-41.5 pp and -30.2 pp) rather than a single threshold crossing. (Welch's t on the binary outcomes, a normal-approximation convenience reported for reference only, gives $t = 2.24$ ($p = 0.025$) and $t = 1.83$ ($p = 0.067$).) Both models fail on the same four philosophical items (P003, P004, P007, P009): the barrier is item-side, not model-side noise.

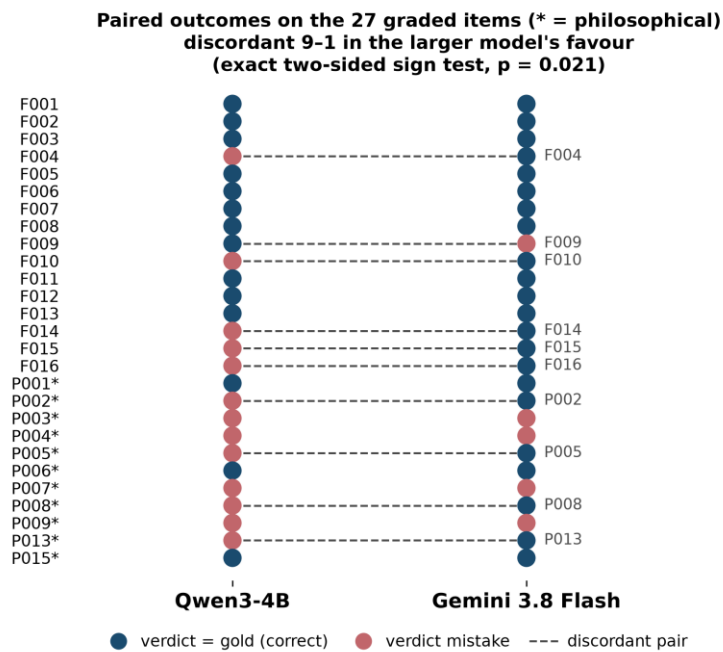


Fig. 1. Paired per-item outcomes on the 27 graded items: the 9–1 discordance in the larger model's favour (exact two-sided sign test, $p = 0.021$).

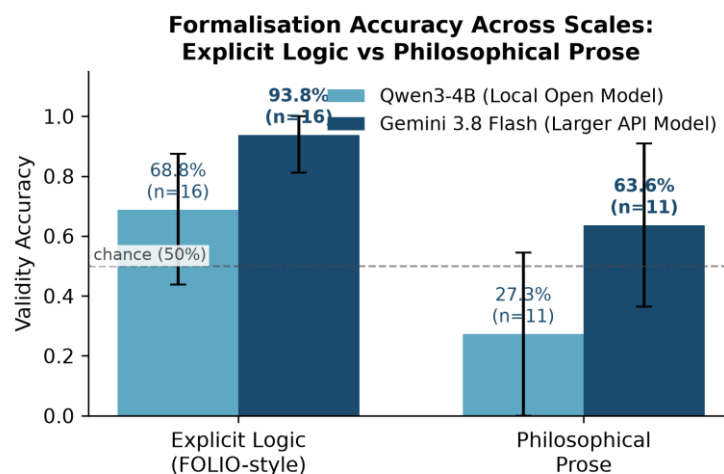


Fig. 2. Validity-verdict accuracy across scales: Explicit logic vs. philosophical prose.

2-4 The Enthymeme Barrier (Suppressed Warrants)

Table 3 and Fig. 3 demonstrate the impact of suppressed warrants. For Qwen3-4B, accuracy on enthymemes was 0.0%. While Gemini 3.8 Flash recovered warrants with explicit textual cues (e.g., al-Ghazali's habit argument, P005), it failed on dialectical enthymemes such as Descartes' dream hypothesis (P007) and Mulla Sadra's modal proof (P003), yielding 50.0% accuracy. A four-variant prompt sweep on a 19-item graded sample tests whether framing is the cause: a warrant-eliciting prompt lifts the larger API model from 7/11 to 10/11 philosophical items (its two recovered enthymemes become four) at a cost on the explicit subset (8/8 → 6/8), while the small model gains only 4/11 → 5/11 (and drops 7/8 → 5/8 on its explicit subset) — recovering a warrant is a genuine competence, unevenly distributed, not a prompt trick either way.

Table 3. Accuracy on Explicit vs. Enthymematic Arguments

Model Architecture	Closed Explicit Premises (n=23)	Enthymematic Arguments (n=4)
Qwen3-4B-Instruct	60.9%	0.0%
Gemini 3.8 Flash	87.0%	50.0%

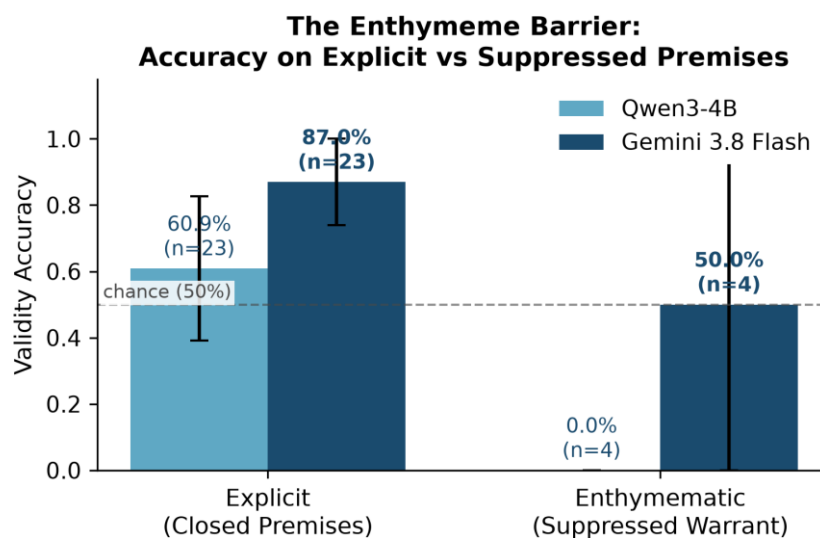


Fig. 3. The Enthymeme Barrier: Accuracy on closed explicit premises versus suppressed warrants.

2-5 Error Taxonomy and Directional Asymmetry

Table 4 summarises the error taxonomy and Fig. 4 the per-item verdict confusion. Notably, 100% of Gemini's verdict mismatches (5/5) and 91% of Qwen's (10/11) followed the directional bias Valid → Invalid. The models do not invent false proofs; they under-specify the premise set, allowing Z3 to construct satisfying countermodels. Across both models, 15 of the 16 mismatches move gold Valid → Invalid; the single exception (F004) marks up a fallacious affirming-the-consequent syllogism as valid. Two denominator conventions must be kept apart.

Table 4 and Fig. 4 operate at verdict level: Table 4 counts all 31 attempted items the four contested items returned a verdict (all invalid) and fall under the Correct rows, although they carry no determinate gold verdict (a generous-basis sensitivity over all 31 appears in Section IV.C) and Fig. 4 counts decided graded items only, so Qwen's two parse errors are excluded and its confusion accuracy is $14/25 = 56.0\%$ rather than the graded $14/27 = 51.9\%$ of Table 1.

Table 4. *Distribution of Verdict-Level Outcomes Across All 31 Attempted Items*

Classification	Qwen3-4B	Gemini 3.8 Flash	Diagnostic Mechanism
Correct (Immediate)	17 (54.8%)	25 (80.6%)	Valid SMT; matching solver verdict
Correct (Repaired)	1 (3.2%)	1 (3.2%)	Repair closed the formula; verdict agrees with annotation
Verdict Mismatch	11 (35.5%)	5 (16.1%)	Well-formed SMT; wrong entailment or uncaptured warrant
Parse / Syntax Error	2 (6.5%)	0 (0.0%)	Solver crash or ill-formed S-expression

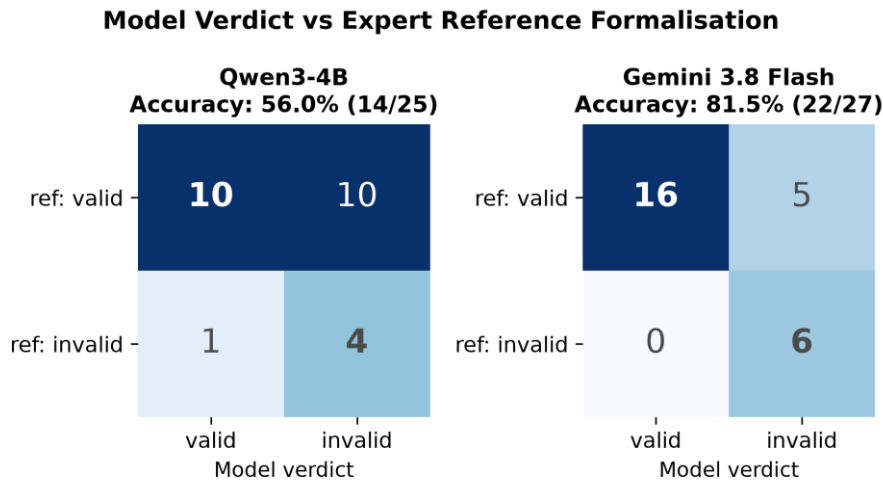


Fig. 4. *Model verdict versus expert reference formalisation (decided graded items only; Qwen 14/25 = 56.0%).*

2-6 Three-Way Verification on FOLIO: A Four-Run, Three-Family Replication

The full 204-item FOLIO validation split was evaluated under the three-way protocol in four runs on the identical file: (i) an archived pilot whose backend alias did not pin model identity (138 attempts lost to provider-side HTTP 503 errors, $n=66$ usable); (ii) a fully pinned Gemini 3.8 Flash re-run (204/204 items, identity pinned by the upstream model field of every response); (iii) a fully pinned Qwen3-4B re-run through a local Ollama runtime (204/204 items, identity pinned by the model-file digest); and (iv) a fully pinned Nemotron-3-Ultra-550B re-run (204/204 items, identity pinned by the upstream model field, run with a 4000-token completion budget and low reasoning effort). Table 5 and Fig. 5 summarise; solver or parse failures (7 in the pilot, 6 Gemini, 99 Qwen, 1 Nemotron) count as failures in the per-class accuracies.

Table 5. *Three-Way FOLIO Validation: Per-Class Accuracy Across Four Runs*

Run (identity)	True	False	Uncertain	Overall
Pilot, n=66 (alias, unpinned)	62.5% (15/24)	25.0% (4/16)	92.3% (24/26)	65.2%
Gemini 3.8 Flash, n=204 (pinned)	72.2% (52/72)	54.0% (34/63)	95.7% (66/69)	74.5%
Qwen3-4B, n=204 (pinned)	27.8% (20/72)	15.9% (10/63)	44.9% (31/69)	29.9%
Nemotron-3-Ultra-550B, n=204 (pinned)	73.6% (53/72)	57.1% (36/63)	97.1% (67/69)	76.5%
Always-Uncertain baseline	0.0% (0/72)	0.0% (0/63)	100.0% (69/69)	33.8%

Three findings follow:

- 1) The asymmetry is not an infrastructure artifact: the pinned Gemini run reproduces the pilot's signature at full scale and agrees with the pilot on all 56 items that received a three-way verdict in both runs (Clopper–Pearson 95% lower bound for the agreement: 93.6%).
- 2) The failure to refute replicates across model families: on items that received an actual verdict, accuracy on Gold-Uncertain items far exceeds accuracy on Gold-False items in all three backends (97.1% vs 56.7% for Gemini; 86.1% vs 28.6% for Qwen; 97.1% vs 58.1% for Nemotron). What replicates is the asymmetry itself, not one abstention profile: on Gold-False items only the small model abstains more often than it refutes (17 vs 10), while the larger-API models refute more than they abstain (Gemini 34 vs 22; Nemotron 36 vs 22)—yet decided-item refutation accuracy never exceeds 58.1% (36/62) in any family; counting the unresolved items as failures the bound is 57.1% (36/63). The small local model additionally loses 48.5% of items to parse errors—but a control re-run with an 8192-token context window still loses 45.1%, so truncation is not the main cause; and an offline ablation shows the symbolic repair layer rescues 23 of Qwen's 105 decided items, so repair is not the cause either. Note also that the trivial Always-Uncertain baseline (33.8%) exceeds the Qwen run's overall accuracy—the per-class columns, not the overall row, carry the finding.
- 3) A direct no-solver baseline on the curated corpus (asking the model for a one-word verdict) splits the models: Qwen3-4B rises to 66.7% (above its 51.9% through the pipeline) and refutes 5/6 gold-False items—for the small model, formalisation breaks refutation; Gemini 3.8 Flash falls to 59.3% and never predicts False at all (0/6 on gold-False)—the larger API model's raw judgement refuses to refute, and the SMT loop is what enables its refutations. A paired exact sign test over the two backends' 204 outcomes confirms the aggregate gap is systematic: 101 vs 10 discordant items in the larger model's favour ($p = 4.4 \times 10^{-20}$); the third pinned family slots beside the larger model (76.5% vs 74.5% overall; 99 vs 4 discordant items against the small model). These are not benchmark claims: the runs use the validation split with our own three-way prompt.

The finding we defend is the asymmetry itself—refutation is a distinct and harder competence than confirmation or abstention—now resting on three independently pinned backends from different model families.

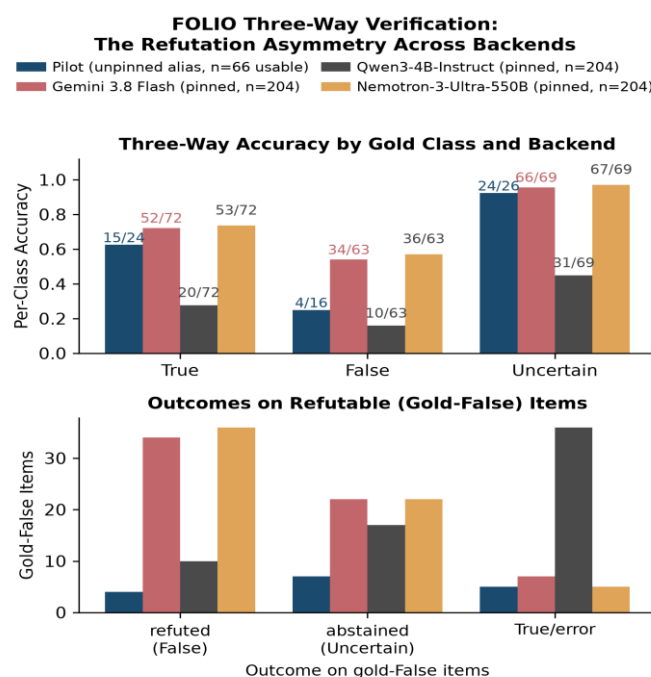


Fig. 5. Three-way FOLIO verification across four runs: per-class accuracy by gold class and backend (left); outcomes on refutable (gold-False) items (bottom).

For orientation only: on the FOLIO test split (226 items, pinned model identities, three-valued labels), the original authors report GPT-4 few-shot at 64.2%, GPT-4 with chain-of-thought at 68.9%, LINC (Olausson et al., 2023) 73.1%, Logic-LM 78.1%, and human expert annotators at 96.0% (Han et al., 2024, Table 4 and Sec. VI-A); the DetermLR system is listed there at 77.5%. These are not directly comparable to our runs (validation split, three-way protocol, our own prompt); the comparable finding is the per-class asymmetry, which an aggregate score cannot show, and which complements Han et al.'s observation that models tend to predict True over False or Unknown (their Appendix J).

3. Discussion and Theoretical Implications

3-1 Mechanical Transcription vs. Hermeneutic Reconstruction

Our results establish a fundamental divergence between transcribing surface syntax and reconstructing an argument. Language models operate as rigid literalists: by failing to supply implicit Waltonian warrants (Walton et al., 2008), the warrant in Toulmin's sense (Toulmin, 2003), they reduce valid philosophical reasoning to formal fallacies. The degradation on prose

echoes the compositional limits of transformers (Dziri et al., 2023) and the answer-distribution biases of pretrained reasoners (Saparov & He, 2023); the direct no-solver baseline of Section III-F relocates the failure in the formalisation step itself. This places the present results in the lineage of formal-argument work in this journal (Nabavi & Mirsanei, 2026), with a tool whose purpose is not to replace interpretive labour but to render it inspectable—the Fregean ideal of letting derivation decide (Frege, 1879), applied where the rendering is the hard part.

3-2 Countermodels as Automated Thought Experiments

Whenever Z3 returns 'sat', it synthesizes a possible world where premises hold and the conclusion fails. For Descartes' dream argument (P007), the archived countermodel is a world in which the external world is perceived by the senses, may be a dream, and is nevertheless certain—possible only because the suppressed warrant 'whatever may be a dream is not certain' never entered the formalisation. The solver does not mis-derive; it displays, as an inspectable object, exactly the underdetermination a careful reader would have diagnosed.

3-3 Expressibility Boundaries in Classical Philosophy

For all 4 contested arguments (Plato, Kant, Aquinas, Ibn Rushd), both models returned 'invalid'. These deductions require modal necessity, deontic operators, epistemic modality, or supremum principles, identifying clear boundaries for first-order SMT solvers in computational humanities. The contested/graded partition is fixed by the corpus annotations and argued in each item's published note (Section III-A), grounded in the modal, deontic or epistemic content the gold text requires; we did not pre-register the decision externally, and first-order approximations of the four items do exist in the archived reference set — they were built for the sensitivity check that follows, so the partition is a judgment call we own, not a proven a-priori constraint. As a sensitivity check, a generous alternative scoring that credits every returned verdict on a contested item as agreement leaves the ranking and the gap unchanged (83.9% vs. 58.1% over all 31 items); it raises the headline numbers, which is why it is a bound, not the metric.

4. Limitations

- Sample size: 15 philosophical arguments (11 graded); exact intervals are wide and the nonparametric tests borderline. The finding's weight rests on the replicated effect size and the paired exact test ($p = 0.021$), not on any single threshold.
- Single expert annotation: the reference formalisation is one expert's reading; there is no second annotator and no inter-annotator agreement statistic. The classical arguments are also studied in English translation (via the editions listed in the corpus), so the benchmark tests agreement with one translator-mediated reading of gold validity produced by the authors themselves — a threat to construct validity that per-item reconstruction confidence is meant to discount, not eliminate. Scale is confounded: the two models differ in family, training data, quantisation and serving channel; we claim an association “larger model $\leftrightarrow$ smaller gap”, not a causal effect of scale.

Contamination cannot be excluded: classical arguments are almost certainly in both models' training data, which may flatter the philosophical column.

- Segmentation is lexical and marker-based; all corpus arguments carry an explicit conclusion marker, so the marker-free fallback path is untested on real data, and the short Persian marker is ambiguous in compound verb phrases.
- The direct no-solver baseline (66.7% for the small model, above the pipeline's 51.9%; 59.3% for the larger model, below its 81.5%) shows the SMT loop's accuracy effect is model-dependent; we report this and argue its value is inspectability and diagnosis, not raw accuracy.

Conclusion and Future Work

This paper presented a neuro-symbolic pipeline combining discourse segmentation, LLM formalisation, automated repair, and Z3 verification. Across 31 benchmark arguments and the full FOLIO validation split, the larger model reaches a 100% parse rate on the curated corpus but leaves a 30–42 percentage-point gap on philosophical prose due to unrecovered warrants, and a refutation asymmetry (accuracy on refutable items lagging far behind accuracy on undeterminable ones) that replicates across three independently pinned model families. Future work will investigate neuro-symbolic abductive search to hypothesize suppressed warrants, extending the framework to modal and intensional logics.

References

- de Moura, L., & Bjørner, N. (2008). Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems* (pp. 337–340). Springer. https://doi.org/10.1007/978-3-540-78800-3_24
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36. <https://arxiv.org/abs/2305.18654>
- Frege, G. (1967). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought (S. Bauer-Mengelberg, Trans.). In J. van Heijenoort (Ed.), *From Frege to Gödel: A source book in mathematical logic, 1879–1931* (pp. 1–82). Harvard University Press. (Original work published 1879)
- Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., Coady, J., Peng, D., Qiao, Y., Benson, L., Sun, L., Wardle-Solano, A., Szabó, H., Zubova, E., Burtell, M., Fan, J., Liu, Y., Wong, B., Sailor, M., ... Radev, D. (2024). FOLIO: Natural language reasoning with first-order logic. In *Proceedings of EMNLP 2024* (pp. 22017–22031). <https://doi.org/10.18653/v1/2024.emnlp-main.1229>
- Nabavi, L., & Mirsanei, S. A. (2026). Paraconsistency, relevance and vagueness: A proof theoretic approach to fuzzy relevant logic FR. *Journal of Philosophical Investigations*, 20(55), 325–352. <https://doi.org/10.22034/jpiut.2026.70950.4391>
- Olausson, T. X., Gu, A., Lipkin, B., Zhang, C., Solar-Lezama, A., Tenenbaum, J. B., & Levy, R. (2023). LINC: A neurosymbolic approach for logical reasoning by combining language models with first-

- order logic provers. In *Proceedings of EMNLP 2023* (pp. 5153–5176). <https://doi.org/10.18653/v1/2023.emnlp-main.313>
- Pan, L., Albalak, A., Wang, X., & Wang, W. Y. (2023). Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of EMNLP 2023* (pp. 3806–3824). <https://doi.org/10.18653/v1/2023.findings-emnlp.248>
- Saparov, A., & He, H. (2023). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *Proceedings of ICLR 2023*. <https://arxiv.org/abs/2210.01240>
- Tafjord, O., Dalvi, B., & Clark, P. (2021). ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL 2021* (pp. 3621–3634).
- Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press. (Original work published 1958)
- Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.
- Ye, X., Chen, Q., Dillig, I., & Durrett, G. (2023). SatLM: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 36. <https://arxiv.org/abs/2308.10457>