pamaldi at SemEval-2026 Task 11: Neuro-Symbolic Syllogistic Reasoning via LLM-Guided Structure Extraction and Deterministic Validation

Pasquale Grimaldi
Università degli Studi di Firenze
pasquale.grimaldi@edu.unifi.it

Abstract

We describe our participation in SemEval-2026 Task 11, Subtask 1: determining the formal validity of syllogisms in English while minimizing the influence of content plausibility. Our system implements a neuro-symbolic pipeline that strictly separates neural and symbolic components. An LLM extracts the formal structure of natural-language syllogisms—proposition types (A, E, I, O) and the three terms—while the syllogistic figure is computed deterministically and a symbolic validator checks whether the resulting mood–figure pair belongs to the 24 classically valid Aristotelian forms. On the official evaluation we achieve 96.34% accuracy, Total Content Effect (TCE) of 1.02, and combined score of 56.57. Compared to pure-LLM baselines on the same backbone, our system more than doubles the combined score (from 26.52 to 56.57) and reduces TCE by nearly an order of magnitude. Swapping the extractor to Claude Sonnet 4.5 preserves combined score and TCE, confirming that content-invariance is contributed by the symbolic stage rather than any particular LLM. A paraphrase probe reveals that the validator is invariant to surface form but the extractor is sensitive to premise ordering—a specific, fixable limitation we identify as the primary target for future work.

1 Introduction

Large language models (LLMs) exhibit a persistent vulnerability to *content effects* in logical reasoning: they tend to accept invalid arguments when the conclusion is plausible, and reject valid arguments when it is implausible (Dasgupta et al., 2022; Eisape et al., 2024). This entanglement of formal reasoning with world knowledge undermines LLM reliability on tasks that require strict logical inference.

SemEval-2026 Task 11 (Valentino et al., 2026) evaluates systems on syllogistic reasoning under four subtasks. We focus on Subtask 1: predicting the formal validity of categorical syllogisms in English, where the ranking metric jointly rewards accuracy and penalizes content bias through the Total Content Effect (TCE).

Our key insight is that syllogistic validity is a purely structural property: it depends entirely on the *mood* (the triple of proposition types) and the *figure* (the arrangement of terms), not on premise truth. We therefore decompose the problem into (1) a *neural extraction* stage that parses the syllogism into formal components and (2) a *symbolic validation* stage that checks membership in the 24 classically valid Aristotelian forms.

Our contributions are: (1) a strict neuro-symbolic decomposition in which the LLM is forbidden from producing reasoning steps and emits only four AEIO labels plus three term strings, with all inference performed by closed-form lookup; (2) a controlled comparison on the same backbone against direct-prompt, chain-of-thought, self-consistent CoT, and an LLM-generates-Prolog variant, showing that reducing the LLM's reasoning role—rather than adding sampling—closes the content-effect gap; (3) a branch-level decomposition localizing all residual content sensitivity to the neural components (the extractor and the fallback), not the validator; (4) a multi-model study showing the design claim holds across backbones; and (5) a paraphrase probe that isolates premise-order sensitivity as the primary residual weakness.

Our submitted system achieves a combined score of 56.57 with 96.34% accuracy and TCE of 1.02. Code and prompts are public.[1,2]

2 Background

Task and metrics. SemEval-2026 Task 11 Subtask 1 (Valentino et al., 2026) requires predicting the binary validity label of English categorical syl-

[1] https://github.com/pamaldi/pamaldi_semeval_task11

[2] Language of data: English.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1436–1442
July 3-4, 2026 ©2026 Association for Computational Linguistics

logisms. The training set contains 200 instances and the test set 191, each balanced across four validity–plausibility conditions: VP, VI, IP, II. Systems are evaluated on overall accuracy, Total Content Effect (TCE, lower is better), and the primary ranking metric

$$\text{Score} = \frac{\text{ACC}}{1 + \ln(1 + \text{TCE})}.$$

Categorical syllogisms. A categorical syllogism has three terms (major P, minor S, middle M) across two premises and a conclusion; each proposition is one of **A**, **E**, **I**, **O**. The *mood* is the ordered triple of types; the *figure* (1–4) is determined by the positions of M across the two premises (Copi et al., 2014). Under the traditional Aristotelian interpretation with existential import, exactly 24 mood–figure pairs are valid.

Related work. Content effects in LLMs are well documented (Dasgupta et al., 2022; Eisape et al., 2024; Bertolazzi et al., 2024; Ozeki et al., 2024; Seals and Shalin, 2024), with mechanistic analyses (Kim et al., 2025). Mitigation work includes activation steering (Valentino et al., 2025; Maraia et al., 2026), quasi-symbolic CoT (Ranaldi et al., 2025), faithful-CoT (Lyu et al., 2023; Xu et al., 2024), and LLM–symbolic theorem-prover hybrids (Quan et al., 2024). Our approach differs in that the LLM is forbidden from producing reasoning: it emits only AEIO labels and term strings, and all inference is closed-form lookup—a stricter separation that yields higher accuracy and lower TCE than an LLM-generates-Prolog variant on the same data (Section 3.1).

3 System

3.1 Design Evolution

The final system emerged from iterative exploration of four architectures, re-executed on the official test set after gold labels were released; results are in Table 1. Early approaches used LLM-generated Prolog with Reflexion self-correction (Shinn et al., 2023) (90.58%/TCE 8.49) and full LLM extraction of types, terms, and figure (97.91%/TCE 2.13). Moving the figure computation from LLM to deterministic lookup yielded the simplified variant (97.38%/TCE 2.08). An incremental meta-learning approach that grew the prompt automatically from training errors overfit (95.29%/TCE 6.25). The **submitted** system combines the most effective elements: simplified extraction, deterministic figure, self-consistency voting, post-processing rules, and LLM fallback, reaching 96.34% accuracy and TCE 1.02—the lowest TCE of all variants by a factor of two or more.

Approach	Acc.	TCE	Score
Prolog + Reflexion	90.58	8.49	27.87
Incr. Meta-Learning	95.29	6.25	31.97
Full LLM Extraction	97.91	2.13	45.74
Simpl. Det. Figure	97.38	2.08	45.81
Final (submitted)	**96.34**	**1.02**	**56.57**

Table 1: Test set (191 instances) across explored approaches.

Figure	Valid Moods
1	AAA, EAE, AII, EIO, AAI, EAO
2	EAE, AEE, EIO, AOO, EAO, AEO
3	AAI, IAI, AII, EAO, OAO, EIO
4	AAI, AEE, IAI, EAO, EIO, AEO

Table 2: The 24 valid syllogistic forms under Aristotelian logic with existential import, by figure.

The central pattern across all variants is that errors originate in neural extraction, never in deterministic validation, and that more deterministic components yield lower content effects.

3.2 Final Pipeline

The submitted pipeline: (1) the LLM extracts proposition types and terms as $k = 3$ samples at temperatures $[0.1, 0.3, 0.5]$, aggregated by majority vote; (2) post-processing rules correct I/O confusion, double negation, and compound-term mismatches; (3) the figure is computed deterministically from the middle term's position; (4) a symbolic validator checks membership in the 24 valid forms (Table 2); (5) on extraction failure (3.7% of test instances), a direct LLM fallback provides the verdict. The prompt template is in Appendix A.

4 Experimental Setup

The extraction and fallback components use Qwen3-32B[3] with few-shot prompting and $k = 3$ self-consistency. To isolate the contribution of the decomposition we compare against three pure-LLM baselines on the same backbone: **B1** direct prompt ("VALID/INVALID", temperature 0); **B2** chain-of-thought ("Think step by step," temperature 0.3); **B3** CoT + self-consistency ($k = 4$ at $[0.0, 0.3, 0.5, 0.7]$, majority vote). For cross-model generalisation we re-run the full pipeline

[3]Accessed via Amazon Bedrock.

System (Qwen3-32B)	Acc.	TCE	Score
B1: Direct prompt	79.58	22.16	19.21
B2: Chain-of-thought	89.01	9.55	26.52
B3: CoT + SC ($k = 4$)	89.53	10.59	25.95
Prolog + Reflexion	90.58	8.49	27.87
Ours (final)	**96.34**	**1.02**	**56.57**

Table 3: Controlled comparison on the test set, same Qwen3-32B backbone. All pure-LLM variants exhibit substantial content effect (TCE ≥ 8); self-consistency sampling (B3) does not close the gap left by CoT (B2).

with Claude Sonnet 4.5 as extractor (Appendix E). The paraphrase probe (Section 5.4) uses Claude Sonnet 4.5 only to generate paraphrases.

5 Results

Our system achieves 96.34% accuracy (184/191), TCE 1.02, combined score 56.57 on the official evaluation. Subgroup accuracy is near-identical across conditions (VP 95.83%, VI 95.83%, IP 97.87%, II 95.83%), yielding the low TCE. The 7 errors comprise 4 false negatives and 3 false positives (distributed VP=2, VI=2, IP=1, II=2).

5.1 Pure-LLM Baselines

Table 3 compares our system against three pure-LLM baselines on the same backbone. Direct prompting exhibits severe bias (TCE 22.16), with the worst subgroup being valid–implausible at 64.58% (Dasgupta et al., 2022; Eisape et al., 2024). CoT roughly halves TCE but shifts rather than eliminates the bias—the weakest subgroup becomes invalid–plausible at 82.98%. Critically, B3 adds $k = 4$ self-consistency on top of CoT and yields *no* TCE improvement (9.55 vs 10.59) despite $4\times$ the compute, confirming that majority voting across samples sharing the same plausibility-driven bias cannot cancel that bias. Closing the content-effect gap requires a structural solution, not more sampling. Since all baselines share our backbone, our $2\times$ improvement in combined score (26.52 $\rightarrow$ 56.57) cannot be attributed to the model's reasoning capacity—it is a consequence of forbidding validity judgments in the LLM.

5.2 Fallback Branch Decomposition

Table 4 decomposes accuracy and TCE by branch. The symbolic path (96.3% of instances) reaches 97.83% accuracy with TCE 2.13; the 7-instance fallback reaches 57.14%. The system-level TCE of 1.02 is an averaging effect across branches with different subgroup compositions rather than a property of either path: the validator is content-invariant by construction, but content sensitivity is re-introduced whenever extraction errors map an instance onto a form whose validity happens to align with plausibility. Reducing TCE therefore requires tightening the neural components (extractor and fallback), since the symbolic validator contributes no content-dependent errors but cannot correct them either.

Path	N	Acc.	TCE	Score
Symbolic only	184	97.83	2.13	45.69
Fallback only	7	57.14	0.00	57.14
System (union)	191	96.34	1.02	56.57

Table 4: Per-branch decomposition. The symbolic validator introduces no errors on correctly extracted structures; residual content sensitivity arises only when extraction errors map instances onto forms whose validity aligns with plausibility. The fallback's TCE of 0.00 reflects its 7-instance subgroup composition, not a structural property.

Extractor	Acc.	TCE	Score	FB%
Qwen3-32B (submitted)	96.34	1.02	56.57	3.7
Claude Sonnet 4.5	97.38	1.06	56.47	1.0

Table 5: Swapping the extractor LLM, with validator, post-processing, and fallback held fixed. Combined score and TCE are stable across backbones; the stronger extractor reduces fallback rate (FB%).

5.3 Multi-Model Extraction

To test whether the architectural claim holds across backbones, we re-run the pipeline with the symbolic components held fixed and swap the extractor (Table 5). Combined score and TCE are essentially unchanged (56.57 vs 56.47; TCE 1.02 vs 1.06); accuracy tracks extractor quality (96.34% vs 97.38%). Content sensitivity remains near-zero regardless of backbone, while the stronger extractor reduces fallback frequency from 3.7% to 1.0%. Three of the five errors made by the Claude-based pipeline are on the same instances that Qwen3-32B also mis-classifies (the Barbara figure case, the E/O-confusion cases), suggesting further TCE gains require stronger extraction or post-processing rather than backbone substitution.

5.4 Paraphrase Probe

We drew a stratified sample of 30 test instances (8 VP, 8 VI, 7 IP, 7 II; seed 42) and generated three

Variant	Acc.	Δ (pp)	Agree.
Original (n=30)	100.00	—	—
Voice (n=30)	93.33	−6.67	93.3
Synonym (n=30)	90.00	−10.00	90.0
Reorder (n=30)	73.33	−26.67	73.3
Pooled (n=90)	85.56	−14.44	85.6

Table 6: Paraphrase probe on 30 stratified instances. Agreement is the fraction of paraphrases on which the pipeline gave the same prediction as on the original. All 13 paraphrase-induced errors are valid→invalid flips, concentrated on the *reorder* variant (8/13), which changes premise order but no words.

paraphrases per instance with Claude Sonnet 4.5 at temperature 0: (a) *voice* (active↔passive), (b) *synonym* (middle-term lexical substitution), (c) *reorder* (premise 2 before premise 1, all text otherwise identical). We then re-ran the full Qwen3-32B pipeline on the 90 paraphrases (Table 6).

The dominant failure mode is *reorder*: 8 of 13 paraphrase-induced errors, despite changing no words. Inspection of the extracted structures confirms the cause: the simplified extractor uses positional cues to assign major/minor-premise roles, so swapping premise order inverts the term-role assignment and the computed figure, mapping valid forms (e.g. AOO-2 Baroco) onto invalid ones (OAO-2). The validator then correctly rejects the (incorrectly) extracted form—a faithful rejection given bad input. This is a specific, fixable limitation: term roles should be derived from the conclusion (S and P appear there, M does not), not from P1/P2 position. All 13 observed errors are valid→invalid flips; none are the reverse, indicating the validator still rejects invalid forms correctly even under extractor stress.

6 Error Analysis and Discussion

All 7 test-set errors are attributable to neural components; the validator introduces none on correctly extracted structures. Four errors occur on the symbolic path (extraction failures upstream of the validator) and three on the fallback. By linguistic trigger: **E/O type confusion** (2/7) where non-standard universal-negative phrasing ("do not overlap"; bare "are not") is parsed as particular O rather than universal E, turning valid Fresison (EIO-4) and Celaront (EAO-1) into invalid OIO-4 and OAO-1; **figure or conclusion-scope error** (2/7), including a failure on Barbara (AAA-1) mis-figured as AAA-3, and a conclusion whose type is genuinely ambiguous in surface form; **compound noun phrase / extraction failure** (3/7), all fallback errors triggered by the middle term appearing in syntactically different forms across premises ("stationery items" vs. "items of stationery"). Per-error detail in Appendix D Table 9. The paraphrase probe adds a fourth mode—premise-order sensitivity—invisible on the canonical test set because originals present P1 before P2 naturally.

Discussion. The symbolic validator is provably content-invariant: Section 5.2 shows it contributes no content-dependent errors, so all residual TCE comes from neural components. The B2→B3 comparison shows adding self-consistency to CoT yields no TCE improvement; structural intervention is required, not compute. Cross-model results (Section 5.3) support that content-invariance is an architectural property rather than a model property: TCE is invariant to backbone, and the stronger extractor mostly reduces fallback invocation. The paraphrase probe identifies premise-order sensitivity as the primary residual weakness; closing this single gap would, all else equal, move paraphrase accuracy from 85.56% to approximately 94%. The decomposition is form-agnostic: extending beyond categorical syllogisms requires swapping the validator (e.g. a SAT/SMT backend) while preserving the extract-then-validate separation.

During development we identified 5 annotation inconsistencies in the training set where universally valid Aristotelian forms (Celarent EAE-1, Baroco AOO-2) were labelled as invalid. These were communicated to the organizers and do not appear in the test set. Because every prediction is explainable by a mood–figure pair, the neuro-symbolic paradigm makes labelling anomalies unusually easy to detect.

7 Conclusion

Our neuro-symbolic system achieves 96.34% accuracy, TCE 1.02, and combined score 56.57, demonstrating near-complete immunity to content effects. Controlled comparisons on a shared backbone show that reducing the LLM's reasoning role—not adding sampling or reflexion—is what closes the content-effect gap: pure-LLM baselines plateau at combined scores of 19–27, our system reaches 56.57. Branch-level decomposition localizes all residual TCE to neural components (extractor 4/7, fallback 3/7). Swapping to Claude Sonnet 4.5 preserves combined score and TCE while reduc-

ing fallback rate to 1.0%, supporting that content-invariance is architectural. A paraphrase probe isolates premise-order sensitivity as the largest residual weakness and points to a concrete fix: computing term roles from the conclusion rather than from P1/P2 position. Future work includes this fix, extending the validator to propositional and first-order fragments, and adapting to the multilingual and irrelevant-premise subtasks.

Limitations

Our system is evaluated only on Subtask 1 (English binary classification over categorical syllogisms). It does not address the multilingual setting of Subtask 3, the irrelevant-premise conditions of Subtasks 2 and 4, or non-categorical logical fragments. Symbolic-path accuracy depends on the extractor LLM's ability to parse noun phrases and quantifier scope; paraphrases and double negations remain dominant extraction failure modes. The LLM fallback—invoked on 3.7% of test instances with Qwen3-32B and 1.0% with Claude Sonnet 4.5—is the other neural component that can reintroduce content sensitivity; reducing fallback by improving extraction is the most direct way to lower system-level TCE. The paraphrase probe on a 30-instance stratified subsample reveals the pipeline's predictions are sensitive to surface transformations that should be logically inert, particularly premise reordering (accuracy drops by 26.67 pp despite preserving text word-for-word). All paraphrase-induced errors trace to extraction and figure computation, not the symbolic validator.

Ethical Considerations

This work uses the SemEval-2026 Task 11 dataset, consisting of synthetic English categorical syllogisms with no personally identifying information. The system performs binary formal-validity classification and is not intended for user-facing deployment; we foresee no direct misuse risks. The extraction and fallback components use external LLMs (Qwen3-32B via Amazon Bedrock; Claude Sonnet 4.5 via the Anthropic API for the multi-model comparison and paraphrase generation), whose training data we do not control; any biases in these models may influence extraction on rare linguistic constructions. By design, the symbolic validator is insensitive to such biases, and the branch-level decomposition (Section 5.2) makes residual sensitivity measurable.

Acknowledgments

We thank the SemEval-2026 Task 11 organizers for their prompt clarifications during the evaluation phase, and the anonymous reviewers for suggestions that shaped the camera-ready version, in particular the baseline comparison, the fallback decomposition, the multi-model study, the error-pattern analysis, and the paraphrase probe.

References

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 (EMNLP)*. Association for Computational Linguistics.

Irving M. Copi, Carl Cohen, and Kenneth McMahon. 2014. *Introduction to Logic*, 14 edition. Pearson.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. 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*.

Tiwalayo Eisape, Michael Henry 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: Human Language Technologies (NAACL)*. Association for Computational Linguistics.

Geonhee Kim, Marco Valentino, and André 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, Vienna, Austria. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*. Association for Computational Linguistics.

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.

Kentaro Ozeki, Risako Ando, Takanobu 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*. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through LLM-symbolic theorem proving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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

S. M. Seals and Valerie L. 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 (NAACL)*. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André 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)*, San Diego, California. Association for Computational Linguistics.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

A Extraction Prompt Template

The extraction prompt instructs the LLM to: (1) identify the two premises and the conclusion; (2) for each proposition, determine the type (A, E, I, O), with special attention to the distinction between I ("Some S are P") and O ("Some S are not P"); (3) identify the three terms (S, P, M) by distribution (S in minor premise and conclusion; P in major premise and conclusion; M in both premises but not the conclusion); (4) output the proposition types (e.g. "EIO") and the three terms in a structured format. The figure (1–4) is then computed deterministically from the position of M across the two premises: Fig. 1 M subject in P1, predicate in P2; Fig. 2 M predicate in both; Fig. 3 M subject in both; Fig. 4 M predicate in P1, subject in P2.

B Development Iterations

It.	Change	Acc.	Score
0	Baseline prompt	70%	20
1	Structured extraction	90%	33
2	Self-consistency voting	94%	33.8
3	I/O post-processing	96%	40
4	24 Aristotelian forms	97%	41.7
5	Compound-term handling	97.5%	45
6	Double-negation handling	98%	48

Table 7: Development progression on the training set (200 instances). Values are approximate.

C Training-Set Ablation

Configuration	Acc.	TCE
Full system ($k = 3$)	97.0%	2.78
− self-consistency ($k = 1$)	94.0%	4.95
− I/O post-processing	96.0%	∼4.0
− 24 → 15 valid forms	93.0%	∼5.5
Baseline (initial prompt)	70.0%	>10

Table 8: Ablation on the training set (200 instances). "−" indicates removal; "24 → 15" replaces the Aristotelian set with its existential-import-free reduction.

Restoring the 24 Aristotelian forms and adding $k = 3$ self-consistency yield the largest gains; I/O post-processing provides a modest but consistent improvement.

D Test-Set Error Details

E Multi-Model Extractor Details

The multi-model study (Section 5.3) uses **Qwen3-32B** (submitted, via Amazon Bedrock) and **Claude Sonnet 4.5** (model ID `claude-sonnet-4-5-20250929`, via the Anthropic API). Both configurations use identical prompt templates, self-consistency sampling ($k = 3$ at temperatures $[0.1, 0.3, 0.5]$), post-processing, deterministic figure computation, symbolic validator, and LLM fallback. Only the extractor LLM differs.

#	Type	Sub.	Gold	Pred.	Path	Trigger
1	FN	VP	AAA-1 (Barbara)	AAA-3	sym	figure misid.
2	FN	VI	EIO-4 (Fresison)	OIO-4	sym	E/O confusion
3	FN	VI	EAO-1 (Celaront)	OAO-1	sym	E/O confusion
4	FP	II	AOE-2[†]	AOO-2	sym	conclusion scope
5	FN	VP	–[‡]	extr. fail	fb	compound NP
6	FP	IP	–[‡]	extr. fail	fb	compound NP
7	FP	II	–[‡]	extr. fail	fb	compound NP

Table 9: All 7 test-set errors with dominant linguistic trigger (Qwen3-32B submitted run). FN=false negative; FP=false positive. Path: sym=symbolic, fb=fallback. [†]Conclusion is ambiguous between O (AOO-2, valid) and E (AOE-2, invalid); gold labels the instance invalid. [‡]Fallback cases have no extracted form; gold is valid for #5 and invalid for #6 and #7.