TUCNLP at SemEval-2026 Task 11: Neuro-Symbolic Content Stripping for Debiased Syllogistic Reasoning

Rafael-Dorian Butas, Alex-Mihai Lapusan, Camelia Lemnaru, Rodica Potolea

Department of Computer Science, Technical University of Cluj-Napoca, Cluj-Napoca, Romania
butas.da.rafael@student.utcluj.ro {alex.lapusan,camelia.lemnaru,rodica.potolea}@cs.utcluj.ro

Abstract

In this paper, we present the solution submitted by TUCNLP at SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Large Language Models. The task requires predicting the formal validity of categorical syllogisms while minimizing susceptibility to content-driven biases in English and 11 additional languages. We show that a modestly-sized model (Qwen3-8B) can achieve near-perfect logical reasoning on the English validity-only subtask, and large reductions in content effect on multilingual and premise-retrieval variants, when augmented with a multi-stage neuro-symbolic pipeline: LLM-based content stripping with iterative error correction converts natural language to abstract categorical forms, a classical symbolic parser validates against the twenty-four Aristotelian syllogistic forms, and asymmetric confidence thresholds mediate between symbolic and neural decisions. Across the four subtasks (ST1 to ST4), our system achieves accuracy ranging from 91.1% to 100% and bias-penalized ranking scores ($\mathcal{M}$) from 31.8 to 100.0, with the main bottleneck being overconfident neural predictions that bypass symbolic verification.

1 Introduction

The ability to perform formal deductive reasoning is considered paramount for machine intelligence, yet Large Language Models (LLMs) frequently struggle to separate logical validity from the semantic plausibility of a conclusion. This phenomenon, known as the *content effect* or *belief bias*, mirrors cognitive limitations long observed in human psychology (Evans, 2003; Lampinen et al., 2024). In such scenarios, a model may incorrectly accept a fallacious argument simply because the conclusion aligns with real-world knowledge, or reject a valid deduction because the conclusion is counterintuitive (Eisape et al., 2024; Seals and Shalin, 2024).

SemEval-2026 Task 11 (Valentino et al., 2026) addresses this challenge by requiring systems to disentangle content-driven heuristics from formal reasoning. The task focuses on categorical syllogisms, i.e. logical structures classically consisting of two premises and a conclusion, where the interplay between truth and validity is systematically manipulated to test for deductive consistency.

We combine a classical syllogistic parser with a LoRA fine-tuned Qwen3-8B (8B parameters), built on the insight that *content stripping*, abstracting away semantic content before reasoning, eliminates content bias on validity-only inputs once they are stripped to abstract form, and reduces it sharply on premise-retrieval and multilingual inputs. A multi-stage pipeline handles LLM-based stripping with iterative error correction, symbolic validation, and cascaded neural-symbolic verification with asymmetric confidence thresholds. Compared to zero-shot Qwen3-8B, our system reduces Total Content Effect (TCE), the average accuracy gap between plausibility-congruent and plausibility-incongruent quadrants, from 21.1 to 0.0 percentage points on English and from 31.4 to 1.1 percentage points on multilingual data, while improving accuracy by up to 37 percentage points. We frame these results as evidence that *content-stripped abstract forms* act as an unbiased proxy: the same mechanism cuts content effect on the out-of-domain NeuBAROCO benchmark without retraining. Our system achieves a perfect score on ST1 (shared by 11/45 teams) and ranks 5th–7th on the remaining three subtasks. Our code is publicly available.[1]

2 Background and Related Work

Syllogistic reasoning has roots in classical logic (Łukasiewicz, 1951) and natural logic traditions (van Benthem, 1986; Sánchez-Valencia, 1991). Recent research focuses on LLMs' systematic con-

[1] https://github.com/ButasRafael/semeval2026-task11

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3408–3421

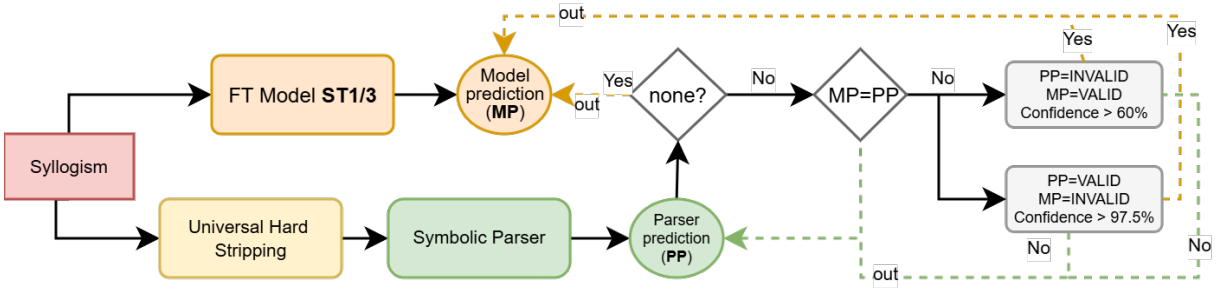

Figure 1: Selective Mode inference (ST1/3). The Symbolic Parser and FT model (FTM) run in parallel producing independent predictions PP and MP, respectively; asymmetric confidence thresholds resolve disagreements.

flation of content plausibility with formal validity. Dasgupta et al. (2022) first demonstrated that LLMs exhibit *content effect*, Lampinen et al. (2024) confirmed that models reason less reliably when semantic content opposes correct inference, Eisape et al. (2024) showed that even CoT-prompted models rarely recognize invalid syllogisms, and Kim et al. (2025) traced these failures to specific reasoning circuits contaminated by world-knowledge attention heads. Benchmarks targeting this problem include SylloBase (Wu et al., 2023) for classical moods, NeuBAROCO (Ozeki et al., 2024) for cross-lingual belief-bias, and the systematic study of Bertolazzi et al. (2024) showing that only supervised fine-tuning reliably mitigates content bias (Wysocka et al., 2025; Seals and Shalin, 2024).

CoT prompting (Wei et al., 2022) and self-consistency decoding (Wang et al., 2023) make intermediate steps explicit but remain susceptible to belief bias (Eisape et al., 2024; Bertolazzi et al., 2024). Quasi-symbolic abstractions (Ranaldi et al., 2025) and fully symbolic CoT methods (Xu et al., 2024; Lyu et al., 2023) translate to formal logic at the cost of formalization overhead. Hybrid neural-symbolic methods include Logic-LM (Pan et al., 2023), which delegates deduction to external solvers; Quan et al. (2024) verify and refine natural-language explanations through LLM-symbolic theorem proving; and PEIRCE (Quan et al., 2025) unifies material and formal inference through iterative refinement. Our system builds on these foundations by using LLM-based content stripping to bridge the formalization gap, converting natural language syllogisms into abstract categorical forms for classical validation, with a fine-tuned neural fallback for unresolvable cases.

3 System Overview

Our system is a hybrid neuro-symbolic architecture that combines a classical syllogistic parser with a LoRA (Hu et al., 2022) fine-tuned Qwen3-8B (Qwen Team, 2025). The central design insight is that *content stripping*, i.e., abstracting away semantic content before reasoning, directly diminishes the content effect. The symbolic component, a classical syllogistic parser backed by Z3 SMT (De Moura and Bjørner, 2008) verification, reasons over abstract logical form with zero content bias, while the neural component handles cases the parser cannot resolve.

3.1 Data Preparation

Content Stripping. We apply LLM-based content stripping using zero-shot Qwen3-8B with in-context examples (8 demonstrations for hard stripping; see Appendix B) in two modes: *soft stripping* replaces content terms with variables A, B, C while preserving grammar, and *hard stripping* normalizes each sentence to one of four categorical forms (*All/No/Some/Some... not X are Y*). A regex based post-processing pipeline handles set-theoretic expressions, double negation, modals, and quantifier variants. When stripping fails (i.e., the output does not parse into exactly three well-formed categorical sentences), up to 3 retries with targeted error feedback are attempted: the system diagnoses the specific failure mode (incorrect sentence count, missing A/B/C variables, missing conclusion marker, or invalid syllogistic figure) and generates corrective hints for each retry, with a thinking-mode fallback on the final attempt.

Variant Generation and Augmentation. Each syllogism is expanded into three variants (original, soft-stripped, hard-stripped) for symbolic verification and training augmentation. For ST2/4, we

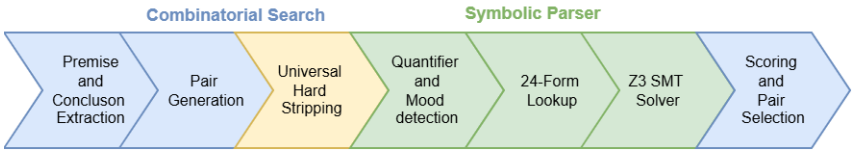

Figure 2: Combinatorial premise search and symbolic parsing pipeline (ST2/4). Premise pairs are hard-stripped, validated against the 24 syllogistic forms (with Z3 fallback), and ranked by a multi-factor scoring function.

add 3–5 LLM-generated filler premises (semantically related, logically irrelevant) in indexed format `[0]...[N]` to train premise selection. For ST3/4, we translate into 11 target languages via DeepL[2] and Google Translate[3], preserving quantifiers and conclusion markers.

3.2 Fine-Tuning

We fine-tune Qwen3-8B using LoRA (Hu et al., 2022) with DoRA (Liu et al., 2024) applied to all attention and FFN projection modules. We train four specialized adapters, one per subtask (FTM-ST1 to FTM-ST4): validity-only adapters for ST1 and ST3 use a lower LoRA rank, while the joint validity and premise retrieval adapters for ST2 and ST4 use a higher rank with RSLoRA (Kalajdzievski, 2023) to accommodate the increased task complexity. A key training design is **quadrant-balanced sampling**: each batch contains equal representation from all four validity×plausibility quadrants (VP, VI, IP, II), directly minimizing TCE during training. For the multilingual subtasks (ST3/4), we extend this with joint quadrant-language stratification. All hyperparameter values are provided in Appendix A.

3.3 Inference: Selective Mode (ST 1 & 3)

For binary validity classification, we use a **parser-primary** architecture with asymmetric neural fallback (Figure 1). Each input is processed along two parallel paths:

(1) Symbolic path. The syllogism is stripped to hard categorical form (for ST3, a *universal stripper* directly converts any language to English categorical form via multilingual LLM prompting) and is then passed to the Symbolic Parser. For cases where the standard lookup fails, a Z3 SMT solver (De Moura and Bjørner, 2008) provides first-order logic verification with existential import (implied in Aristotelian logic, which the task at hand is based on).

(2) Neural path. The fine-tuned model (FTM) predicts validity and extracts a *token-level softmax confidence*, i.e. the softmax probability at the generated validity token position, summed over all tokenizer variants of each label.

Decision logic. If path (1) returns `None` (parse failure), we fall back to path (2). If both agree, we use the parser result. On disagreement, we apply *asymmetric confidence thresholds*, calibrated on the development split: since the FTM is only 45% accurate when contradicting a parser VALID judgment, override to INVALID requires conf ≥ 0.975; conversely, the FTM is 100% accurate when contradicting parser INVALID, so override to VALID requires only conf ≥ 0.60. Below these thresholds, the parser decision stands.

Threshold calibration and robustness. All confidence thresholds (the asymmetric 0.975/0.60 pair above and those listed in Appendix A) are calibrated on the held-out 15% development split using the official metric $\mathcal{M}$; the 45%/100% asymmetry reflects the dev-split conditional accuracy of the FTM when it contradicts a parser VALID vs. INVALID judgment. As a robustness proxy on the test sets, we count parser-FTM verdict disagreements (the only events that can change a prediction): **2/191 on ST1, 7/99 on ST2, 7/190 on ST3, and 26/109 on ST4** (cascaded subtasks counted over parser-ran subset). On ST1 the thresholds are invoked only twice and resolve in favor of the parser in both cases, both correctly, so the ST1 score is robust to any reasonable choice. A formal multi-grid sensitivity sweep is left for future work.

3.4 Inference: Cascaded Verifier (ST 2 & 4)

For joint validity and premise retrieval, we use a three-stage **cascaded verifier** (Figure 3). Because distractor premises share the domain of true premises, content stripping can collapse near-synonymous terms onto the same variable, producing spurious symbolic matches; we therefore adopt an FT-primary design where the symbolic pipeline primarily identifies correct premises rather than classifying validity:

[2] https://www.deepl.com
[3] https://translate.google.com

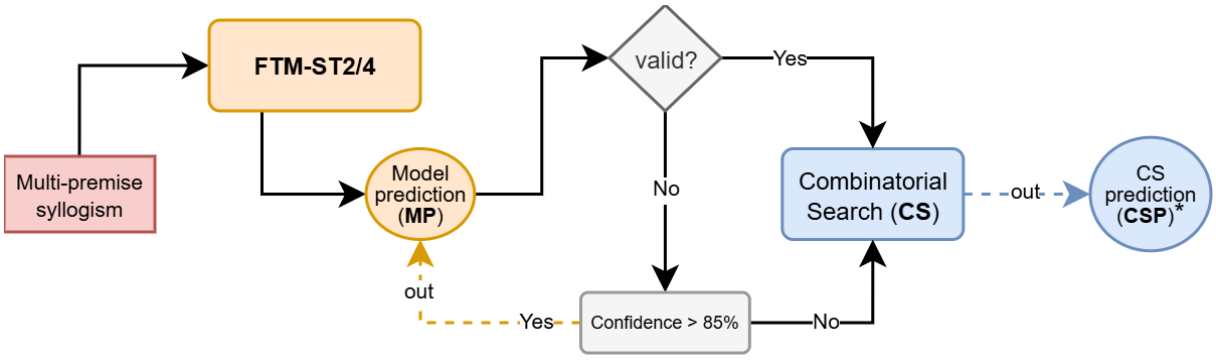

Figure 3: Cascaded Verifier inference (ST2/4). MP = FTM prediction; CSP = combinatorial search prediction. *ST4 adds semantic and per-language guards.

Stage 1: FTM prediction. The fine-tuned model outputs validity, relevant premise indices, and confidence. **Stage 2: Symbolic verification.** If the model predicts INVALID with high confidence (threshold in Appendix A), we return immediately. Otherwise, combinatorial symbolic search generates all $\binom{N}{2}$ premise pairs from the N available premises (Figure 2). A *connectivity gate* pre-filters pairs lacking shared terms via lemmatized overlap (relaxed for non-Latin scripts where lemmatization is unreliable). Each pair is stripped to hard categorical form and checked against the 24 valid forms (with Z3 fallback). When multiple valid pairs are found, a scoring function ranks them by method quality (deterministic match vs. Z3), figure determinacy, FT hint agreement, and for ST4, conclusion alignment via BGE-M3 (Chen et al., 2024).

For ST2, if a valid pair is found, we override to VALID with the symbolic premises. For ST4, a *semantic guard* using BGE-M3 hybrid scoring (dense + ColBERT) must additionally confirm sufficient conclusion term coverage before accepting the override.

Stage 3: Valid premise verification. If the model predicts VALID, we verify the selected premises via symbolic search. For ST4, per-language confidence thresholds select whichever premise pair achieves better conclusion term coverage. A *double-negative safety check* overrides to INVALID if both premises are negative.

4 Experimental Setup

Data. The organizers provide 960 English syllogisms balanced across four validity×plausibility quadrants (Valentino et al., 2026). Training data is English-only; multilingual test data covers 11 additional languages. We apply an 85/15 stratified train/validation split for fine-tuning, preserving quadrant balance. After the augmentation described in §1, ST1 trains on the original English split (809 examples); ST2 additionally incorporates distractor premises (811 training examples); ST3/4 expands to 12 languages (9,792 for ST3; 9,732 for ST4). The test sets contain 191 (ST1), 190 (ST2), and 192 instances each (ST3/4, approximately 17–18 per language, in 11 target languages). Fine-tuning hyperparameters are listed in Appendix Table 4.

Evaluation. Following the official protocol (Valentino et al., 2026), all subtasks are ranked by $\mathcal{M} = S/(1 + \ln(1 + \text{TCE}))$, where S is accuracy for ST1/3 and $\text{Avg}(\text{ACC}, \text{F1}_{\text{premises}})$ for ST2/4. For ST3/4, TCE additionally includes a cross-lingual stability penalty. To isolate the contribution of our pipeline, we compare against the same Qwen3-8B model in a zero-shot setting: identical prompt format but without fine-tuning, content stripping, or symbolic verification.

5 Results

5.1 Main Results

Table 1 presents the performance obtained on the official test set. Our full system achieves perfect accuracy (100.0%), with zero content effect on ST1, and near-perfect results on ST3 (98.4% on the 11-language test set). Compared to the same model in zero-shot mode, the ranking metric $\mathcal{M}$ improves by factors of 5.2× (ST1), 3.2× (ST2), 4.2× (ST3), and 1.8× (ST4). On the official leaderboard, the system achieves the maximum possible score on ST1 (shared by 11 of 45 teams) and ranks 5th/16 on ST2, 6th/15 on ST3, and 7th/15 on ST4.

The results reveal a clear difficulty gradient: English validity-only (ST1) is fully solved; adding either multilinguality (ST3) or premise retrieval

Subtask	System	ACC	$F1_p$	TCE↓	$\mathcal{M}$
ST1	Zero-shot	78.0	–	21.1	19.0
	Ours	**100.0**	–	**0.0**	**100.0**
ST2	Zero-shot	78.6	56.3	18.8	16.9
	Ours	**96.9**	**95.8**	**1.2**	**54.2**
ST3	Zero-shot	60.9	–	31.4	13.6
	Ours	**98.4**	–	**1.1**	**56.7**
ST4	Zero-shot	76.0	52.3	11.6	18.2
	Ours	**91.1**	**90.1**	**5.3**	**31.8**

Table 1: Official test results. Zero-shot: Qwen3-8B without fine-tuning, stripping, or symbolic verification. $F1_p$: premise retrieval F1. $\mathcal{M}$: official ranking metric (§4). Best in **bold**.

(ST2) introduces small residual errors; combining both (ST4) yields the most. Across all subtasks, TCE drops dramatically, from 21.1 to 0.0 (ST1) and from 31.4 to 1.1 (ST3), confirming that content stripping and symbolic verification reduce content bias by an order of magnitude. The component-level ablation (Appendix F, Table 11) shows that fine-tuning and symbolic verification fail on different examples, and that the hybrid pipeline dominates either single component on every subtask.

Out-of-domain generalization. On the independent NeuBAROCO benchmark (Ozeki et al., 2024) of 899 English syllogisms (full numbers in Appendix E), our 8B pipeline reaches $\mathcal{M}$=32.4, beating Gemini-with-thinking (31.8), GPT-5 (23.6), and Claude (24.6) under identical zero-shot prompting and trailing only zero-shot Gemini (35.1) despite a much smaller base model, evidence that using content-stripped abstract forms results in a robust system.

5.2 Content Effect Analysis

Table 2 breaks down the accuracy by the validity×plausibility quadrant. The zero-shot model exhibits the classic content effect pattern: accuracy is highest on Invalid+Implausible instances (where content and logic align in rejecting the argument) and lowest on Valid+Implausible (where the conclusion is counter-intuitive but logically valid). On ST1, the zero-shot gap between these two quadrants is 35.4 percentage points (95.8% vs. 60.4%). Our full system eliminates this asymmetry on ST1: all four quadrants reach 100.0%.

On the multilingual ST3 the zero-shot gap is even larger (92.0% vs. 29.2%, 62.8 points); our system reduces it to 2.2 points, with the residual driven by 3 errors in Swahili (2) and Russian (1). The ST4

	ST1 (EN)		ST3 (ML)	
Quadrant	**ZS**	**Full**	**ZS**	**Full**
Valid + Plaus.	81.2	100.0	50.0	97.9
Valid + Implaus.	60.4	100.0	29.2	97.9
Invalid + Plaus.	74.5	100.0	71.7	97.8
Invalid + Implaus.	95.8	100.0	92.0	100.0
TCE	21.1	0.0	31.4	1.1

Table 2: Per-quadrant accuracy (%) for zero-shot (ZS) vs. our full system on ST1 and ST3.

residual TCE of 5.3 is partly inflated by annotation inconsistencies (§5.3) and partly reflects the added difficulty of premise selection from distractor-rich inputs (per-language results in Appendix C).

5.3 Error Analysis

Across all four subtasks, the system produces 24 apparent validity errors and 9 premise-selection errors against the official ground truth (breakdown in Appendix D). On ST1, all 191 instances are classified correctly: parser and FT agree on every prediction (only 2 cases require Z3 fallback), and the standalone Symbolic-only ablation in Table 11 likewise reaches 100.0 / 0.0. Per-language stripping rates are reported in Appendix C, Table 8. Manual verification against the 24 classically valid Aristotelian forms (assuming existential import) suggests that **12 of the 24 apparent validity errors reflect annotation inconsistencies** in the test set rather than system failures, leaving 12 genuine system errors. The symbolic parser applies the 24-form lookup deterministically, so its verdicts on the inconsistent items reflect the classical reading; the FT-only ablation shows no accuracy drop in the quadrants where these inconsistencies concentrate, indicating analogous training-set noise did not bias the FT model.

Of the 12 identified inconsistencies (detailed in Appendix G), 10 are syllogisms labeled invalid that instantiate classically valid Aristotelian forms (7 unconditionally valid, 3 requiring existential import), and 2 are labeled valid but are in fact invalid. These concentrate in ST4 (10 of 12), reflecting difficulties pertaining to multilingual annotation.

Among the 12 genuine validity errors, the dominant type is **high-confidence FTM false negatives** (7/12, spanning multiple languages and both plausibility conditions, indicating content-independent reasoning failures rather than bias-driven mistakes). The remaining 5 are false positives: 3 from incorrect stripping producing a spurious valid form, 2

	OK	P>F	F>P	Bad	n/a
ST1	189	2	0	0	0
ST2	90	2	5	2	91
ST3	180	1	6	3	2
ST4	74	8	18	9	83

Table 3: Component agreement matrix on the official test predictions. *OK*: both correct; *P>F*: parser correct, FT wrong; *F>P*: FT correct, parser wrong; *Bad*: both wrong; *n/a*: parser not invoked (ST3 also counts 2 parse failures). FT-saves-Parser dominates on ST2/3/4 once stripping is involved.

from FTM overrides of a correct symbolic rejection. No single language dominates the ST4 error distribution, and stripping accounts for only 25% of genuine errors.

5.4 Parser–FTM Synergy

Cross-classifying every test prediction by parser and FT verdict against ground truth (Table 3; details in Appendix H) surfaces three patterns. **ST1 is degenerate**: parser and FTM disagree only twice, both resolved correctly by the asymmetric thresholds. **On ST2/ST3/ST4 the FT corrects more parser errors than the reverse** (F>P exceeds P>F by 3, 5, and 10 instances), refuting the intuition that the symbolic component is uniformly more accurate once stripping is involved. **ST4 shows the strongest two-way complementarity** (8 P>F and 18 F>P); the 9 Both-wrong cases overlap heavily with the annotation inconsistencies of §5.3, with both components reaching the classical Aristotelian verdict that the mislabelled ground truth contradicts.

6 Conclusion

We introduce a hybrid neuro-symbolic system for SemEval-2026 Task 11 that combines a LoRA fine-tuned Qwen3-8B with a classical syllogistic parser operating on content-stripped logical forms. By decomposing inference into a multi-stage pipeline with iterative error correction, asymmetric confidence thresholds, and classical logic safety checks, the system achieves perfect accuracy on English ST1 (shared by 11/45 teams) and ranks 5th–7th on ST2–ST4. Our results demonstrate that a modestly-sized 8B model, augmented with symbolic reasoning over content-stripped forms, can largely eliminate the content bias that persists even in much larger models under zero-shot prompting. More broadly, content-stripped abstract forms behave as a debiasing primitive: the same mechanism transfers to NeuBAROCO without retraining (§E), and we conjecture it applies to any reasoning task whose validity is determined by a closed formal calculus. Among genuine errors, the primary bottleneck is overconfident neural predictions (58%) that bypass symbolic verification.

The symbolic parser guarantees a correct output provided that the parsing was done without any mistakes. One could also argue that the problem of content stripping is more tractable than direct syllogistic reasoning in the context of LLMs; this leads us to believe that major improvements could be achieved through training a perfected parsing model, work that falls into the "future development" category. Improving FTM calibration and applying activation steering (Valentino et al., 2025; Maraia et al., 2026) are also promising directions for further gains.

Limitations

Our approach has several limitations. The Symbolic Parser handles only classical categorical syllogisms (two premises, three terms, four quantifier types) and cannot extend to multi-step reasoning or non-standard quantifiers. Content stripping depends on LLM-based formalization, which is difficult to debug for low-resource and morphologically complex languages (Swahili, Bengali, Telugu). The test sets contain only 17 to 18 instances per language, making per-language estimates statistically noisy. An ablation isolating quadrant-balanced sampling from fine-tuning, and a formal multi-grid sweep over the confidence thresholds, are left for future work. Finally, our identification of 12 annotation inconsistencies (§5.3) assumes Aristotelian logic with existential import; under modern Boolean semantics, 3 of these (Darapti and subalternation) would be reclassified as genuine errors.

Acknowledgments

We express our sincere gratitude to the organizers of SemEval-2026 Task 11 for designing this challenging shared task. This work is supported by the project “Romanian Hub for Artificial Intelligence-HRIA”, Smart Growth, Digitization and Financial Instruments Program, MySMIS no. 351416 and the Technical University of Cluj-Napoca.

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*, pages 13882–13905. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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 *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer.

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

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459.

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

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*.

Geonhee Kim, Marco Valentino, and Andre 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. Association for Computational Linguistics.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*.

Jan Łukasiewicz. 1951. *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*. Oxford University Press.

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 (Volume 1: Long Papers)*, pages 305–329. Association for Computational Linguistics.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and André Freitas. 2026. Abstract activation spaces for content-invariant reasoning in large language models. *arXiv 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*, pages 16063–16077. Association for Computational Linguistics.

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

Xin Quan, Marco Valentino, Danilo Carvalho, Dhairya Dalal, and Andre Freitas. 2025. PEIRCE: Unifying material and formal reasoning via LLM-driven neuro-symbolic refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 11–21. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre 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*, pages 2933–2958. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the*

63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17222–17240. Association for Computational Linguistics.

Victor Sánchez-Valencia. 1991. *Studies on Natural Logic and Categorial Grammar*. University of Amsterdam.

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

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)*.

Johan van Benthem. 1986. *Essays in Logical Semantics*. Springer.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Yongkang Wu, Meng Han, Yutao Zhu, Lei Li, Xinyu Zhang, Ruofei Lai, Xiaoguang Li, Yuanhang Ren, Zhicheng Dou, and Zhao Cao. 2023. Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2347–2367. Association for Computational Linguistics.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and André 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: Human Language Technologies (Volume 1: Long Papers)*, pages 7235–7258. 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

A Implementation Details

Table 4 lists the per-subtask hyperparameters for all four adapters.

Hyperparameter	FM1	FM2	FM3	FM4
LoRA rank r	16	32	16	32
LoRA α	32	64	32	64
RSLoRA	–	Yes	–	Yes
Epochs	4	4	2	2
Batch / GPU	4	4	4	4
Grad. accum.	8	8	8	8
Eff. batch size	32	32	32	32
Learning rate	1e-4	1e-4	5e-5	5e-5
Max seq. length	1024	1024	1024	1536
Training examples	809	811	9,792	9,732
GPUs (V100-32GB)	2	4	2	4

Table 4: Per-subtask hyperparameters. All adapters use Qwen3-8B with DoRA, LoRA dropout 0.1, AdamW optimizer, cosine LR schedule with 10% warmup, weight decay 0.1, and FP16 mixed precision. Early stopping monitors the primary metric.

Training. All adapters are trained with AdamW, cosine learning rate schedule with 10% warmup, and FP16 mixed precision on 2 to 4 NVIDIA Tesla V100-SXM2-32GB GPUs using PyTorch DDP.[4] English subtasks (ST1/2) train for 4 epochs with learning rate 10^{-4}; multilingual subtasks (ST3/4) train for 2 epochs at 5×10^{-5} over the larger translated corpus. Early stopping monitors the official ranking metric on the held-out evaluation split. Training ranges from approximately 2 hours (ST1) to 6 hours (ST4).

Inference thresholds. All confidence thresholds are calibrated on the development split. For the cascaded verifier (ST2/4): the high-confidence INVALID bypass threshold is 0.85; the BGE-M3 semantic guard requires ≥ 0.93 relative conclusion term coverage; and per-language FT confidence thresholds range from 0.80 (Bengali, Telugu) to 0.95 (high-resource Latin-script languages), with intermediate values for Russian (0.93), Chinese (0.85), and Swahili (0.88).

[4] PyTorch 2.5, HuggingFace Transformers 4.57, PEFT 0.18, Accelerate 1.12, Z3 4.15, and BGE-M3 embeddings.

B Data Preparation Examples

Table 5 illustrates the two content stripping modes on syllogisms from opposite content effect quadrants. The first example is *invalid but plausible*: its conclusion aligns with common sense, tempting a model to predict valid. The second is *valid but implausible*: its absurd content tempts a model to reject a logically sound argument. In both cases, hard stripping reduces the input to abstract categorical form, enabling the symbolic parser to determine validity without content interference.

Example 1: Invalid + Plausible (content → false positive)	
Original	All cars are a type of vehicle. No animal is a car. Therefore, no animal can be a vehicle.
Soft	All A are a type of B. No C is an A. Therefore, no C can be a B.
Hard	All A are B. No C are A. Therefore, No C are B.
Example 2: Valid + Implausible (content → false negative)	
Original	Every cat is an invisible creature. A number of cats are animals. Consequently, a portion of animals are invisible.
Soft	Every A is a B. A number of A are C. Consequently, a portion of C are B.
Hard	All A are B. Some A are C. Therefore, Some C are B.

Table 5: Content stripping examples. Soft stripping preserves grammar while replacing content terms with variables; hard stripping normalizes to canonical categorical form.

Table 6 shows two examples of LLM-augmented premise sets used for ST2/4 training. Each syllogism is expanded with 4–5 thematically related but logically irrelevant filler premises (shown in gray). The model must identify the relevant premise pair (bold) and determine validity.

Example 1: Valid, relevant premises = {4, 5}

[0] There exist some vehicles that are neither cars nor motorcycles.
[1] A portion of vehicles are electric.
[2] Some vehicles are used for long-distance travel.
[3] Every vehicle is either a car, a motorcycle, or something else entirely.
[4] Nothing that is a car is a motorcycle.
[5] A few things called vehicles are identified as being cars.
[C] Therefore, some vehicles are not motorcycles.

Example 2: Invalid, relevant premises = {}

[0] Every person in the city has at least one profession.
[1] A portion of professionals work more than one job.
[2] No doctor is a lawyer.
[3] Lawyers are, without exception, professionals.
[4] There exist people who are both engineers and artists.
[5] Some teachers are also part-time counselors.
[6] All doctors have a medical degree.
[C] Therefore, a number of professionals are doctors.

Table 6: LLM-augmented premise sets for ST2/4. Gray: filler premises. Bold: relevant premises (empty set for invalid syllogisms). In Example 2, premises [2] and [3] are the original premises but yield an invalid conclusion (undistributed middle).

Hard Stripping Prompt Structure. The hard stripping prompt instructs the LLM to convert each syllogism into exactly three categorical sentences using variables A, B, C. It specifies the four valid output forms (*All/No/Some/Some... not X are Y*), provides 8 in-context demonstrations covering all four quantifier types, and includes detailed negation-handling rules: (1) *not all/every* → *Some X are not Y*; (2) *not any* → *No X are Y*; (3) double negatives cancel (*no X are not Y* → *All X are Y*); (4) predicate negation (*every X is not Y* → *No X are Y*).

Error Diagnosis on Retry. When stripping fails validation, the system classifies the failure and generates targeted corrective hints before each retry:

- *Sentence count* $\neq 3$: split or merged premises.
- *Missing A/B/C variables*: any of the three logical variables absent from the output.
- *Missing conclusion*: output lacks a "Therefore," marker.
- *Invalid figure*: no shared middle term across premises.

C Additional Results

Table 7 shows the per-language validity accuracy for the multilingual subtasks. On ST3, our system achieves 100% on 9 of 11 languages; only

Swahili (88.2%) and Russian (94.1%) have errors. On ST4, Portuguese is the only error-free language. The worst-performing language against the official ground truth is Russian (76.5% on ST4, 4 of 17 apparent errors), followed by Swahili (82.4%). However, as discussed in §5.3, manual verification suggests that 3 of Russian's 4 ST4 errors are annotation inconsistencies (the syllogisms instantiate valid Barbara or Darapti forms), leaving only 1 genuine system error for Russian. After accounting for these inconsistencies, the per-language error distribution in ST4 is more uniform, with no single language accounting for more than 2 genuine errors.

	ST3		ST4	
Lang	**ZS**	**Full**	**ZS**	**Full**
it	61.1	100.0	66.7	88.9
es	55.6	100.0	77.8	94.4
fr	66.7	100.0	77.8	94.4
zh	61.1	100.0	50.0	94.4
de	66.7	100.0	83.3	94.4
te	82.4	100.0	82.4	94.1
sw	47.1	88.2	88.2	82.4
pt	52.9	100.0	88.2	100.0
nl	64.7	100.0	94.1	94.1
bn	52.9	100.0	58.8	88.2
ru	58.8	94.1	70.6	76.5

Table 7: Per-language validity accuracy (%) on ST3 and ST4. ZS = zero-shot Qwen3-8B; Full = our full system. Each language has 17 to 18 test instances.

Table 8 reports, per language, the rate at which the symbolic component produced a verdict on the official test predictions. ST3 has high parse rates almost everywhere (94–100%); the only failures are 1 Swahili and 1 Chinese instance whose surface forms resisted the four canonical categorical templates. ST4's substantially lower per-language rates (44–72%) do *not* reflect a higher stripping failure rate: by design, the cascaded verifier (§1) invokes the parser only when FT confidence is below the bypass threshold or FT predicts VALID, so the parser is never asked to verdict the easy, FT-confident inputs. The rates therefore describe how often the parser is consulted, not how often it can succeed.

Lang	**ST3 parse-rate**	**ST4 parser-consulted rate**
bn	17/17 = 100.0%	9/17 = 52.9%
de	18/18 = 100.0%	9/18 = 50.0%
es	18/18 = 100.0%	13/18 = 72.2%
fr	18/18 = 100.0%	12/18 = 66.7%
it	18/18 = 100.0%	10/18 = 55.6%
nl	17/17 = 100.0%	8/17 = 47.1%
pt	17/17 = 100.0%	11/17 = 64.7%
ru	17/17 = 100.0%	8/17 = 47.1%
sw	16/17 = 94.1%	11/17 = 64.7%
te	17/17 = 100.0%	10/17 = 58.8%
zh	17/18 = 94.4%	8/18 = 44.4%

Table 8: Per-language rate at which the symbolic component produced a verdict on the official test predictions. ST3 reflects raw stripping success; ST4 reflects the cascaded verifier's invocation pattern (parser is consulted only when FT is uncertain or predicts VALID), not stripping ability per se.

D Error Breakdown

Table 9 categorizes the 33 errors (24 validity + 9 premise-only) across subtasks. ST1 has zero errors and is omitted.

Error Type	**ST2**	**ST3**	**ST4**	**Total**
GT annotation error	2	0	10	12
Strip-induced FP	0	1	2	3
FT-override FP	0	0	2	2
High-conf. FT FN	2	2	3	7
Genuine validity errors	2	3	7	12
Premise-only	2	–	7	9

Table 9: Error categorization across subtasks (ST1 has 0 errors). Top row: annotation inconsistencies where our predictions match valid Aristotelian forms.

E Out-of-Domain Evaluation: NeuBAROCO

To assess generalization beyond the shared task data, we evaluate our Selective Mode pipeline (ST1 adapter with Qwen3-8B) on NeuBAROCO (Ozeki et al., 2024), an independent benchmark of 899 English syllogisms designed to probe belief-bias effects. We compare against three frontier models in zero-shot mode (no fine-tuning, no symbolic verification): GPT-5, Claude, and Gemini (with and without thinking mode). All models receive identical prompting.

System	ACC	TCE↓	$\mathcal{M}$
GPT-5 (zero-shot)	84.97	12.55	23.56
Claude (zero-shot)	89.21	12.80	24.61
Gemini (zero-shot)	92.44	4.13	35.07
Gemini + thinking	95.00	6.29	31.81
Ours (Qwen3-8B + pipeline)	**91.77**	**5.27**	**32.36**

Table 10: Out-of-domain evaluation on NeuBAROCO (899 English syllogisms). Frontier models are evaluated zero-shot under identical prompting (no fine-tuning, no symbolic verification). $\mathcal{M} = \text{ACC}/(1+\ln(1+\text{TCE}))$. At 8B parameters our pipeline beats Gemini-with-thinking, GPT-5, and Claude on $\mathcal{M}$ and trails only zero-shot Gemini, despite using a much smaller base model. Numbers reflect an earlier iteration of the pipeline; the current Selective-Mode adapter is expected to perform at least as well.

Zero-shot Gemini achieves a higher combined score (35.1 vs. our 32.4) at presumably much higher parameter count, while Gemini-with-thinking (31.8) trades a 2.6-point accuracy gain for 2.2 points of additional TCE, confirming that raw scale alone does not eliminate belief bias. Against the other two frontier baselines, our 8B pipeline wins decisively: GPT-5 (23.6) and Claude (24.6) lose roughly 9 points on $\mathcal{M}$ to their content effect (TCE 12.6 and 12.8 respectively, against our 5.3). The fact that the same content-stripping mechanism reduces TCE on inputs the system was never fine-tuned for is the clearest evidence that the debiasing is structural: driven by reasoning over abstract categorical forms rather than memorized from the SemEval-2026 training distribution.

F Ablation Study

Table 11 isolates the contribution of each pipeline component by evaluating, on the official test predictions, what each component would have produced on its own: (1) *Zero-shot*: Qwen3-8B with no fine-tuning or symbolic verification; (2) *FT-only*: the fine-tuned model's verdict in isolation (`ft_validity` / `ft_premises` from the cascaded predictions; `llm_validity` for ST1; `_ft_validity` for ST3); (3) *Symbolic-only*: the symbolic component's verdict in isolation (`symbolic.validity` / `symbolic.premises`; `symbolic_validity` for ST1; `_parser_validity` for ST3); (4) *FT-combin.*: FT model performing combinatorial premise-pair search without the symbolic parser (ST4 only). All rows are scored with the official evaluation script against the same ground truth used for Table 1.

The results show clear complementarity between the neural and symbolic components. Fine-tuning alone reaches 99.0/2.1 on ST1, 97.9/2.2 on ST3, and 89.1/5.2 on ST4 (ACC/TCE), a large improvement over zero-shot but with a residual content effect that the FT model does not eliminate. The symbolic component, when its verdict is used, is essentially bias-free on validity-only inputs (100.0/0.0 on ST1; 95.3/5.2 on ST3 over the 190 parseable items), confirming that reasoning over content-stripped abstract forms removes most of the bias by construction. On ST2/ST4 the parser runs only on FT-uncertain inputs, so its standalone TCE on that biased subset is high (23.9 and 32.2); this is exactly why the cascaded verifier (§1) is FT-primary on premise-retrieval subtasks. The full pipeline integrates both components and inherits the parser's deterministic correctness on cleanly stripped inputs while falling back to the calibrated FT model where stripping fails, dominating either single component on every subtask. **Decomposing the TCE reduction:** on ST1, balanced fine-tuning closes the gap from zero-shot 21.1 down to 2.1 (about 90% of the total reduction), and the symbolic component closes the residual last mile (2.1→0.0); on ST3 the same decomposition is 31.4→2.2 (FT-only contributes 93%) → 1.1 (full pipeline). The symbolic component therefore contributes a smaller absolute amount of TCE reduction, but provides the auditable, content-independent verdict that the FT model alone cannot guarantee, and on ST2/ST4 it additionally constrains premise selection through combinatorial search (full-pipeline F1 jumps from 86.6 to 95.8 on ST2, and from 84.9 to 90.1 on ST4 over their respective FT-only baselines). Due to the relatively small per-language test sample sizes, some differences may appear modest, but they are consistent across configurations.

G Annotation Inconsistency Details

Table 12 lists the 12 test instances identified as annotation inconsistencies in §5.3. For each, we show the ground-truth label, the Aristotelian form instantiated by the relevant premises and conclusion (abstracted to variables A, B, C, D), and the subtask/language. Forms marked with † require existential import, standard in Aristotelian logic. $\not\Rightarrow$ denotes invalid inference.

False negatives. The 10 entries below show the full English surface form of valid syllogisms mis-

	ST1		ST2			ST3		ST4		
Config	**ACC**	**TCE↓**	**ACC**	**$F1_p$**	**TCE↓**	**ACC**	**TCE↓**	**ACC**	**$F1_p$**	**TCE↓**
Zero-shot	78.0	21.1	78.6	56.8	18.8	60.9	31.4	76.0	52.3	11.6
FT-only	99.0	2.1	96.8	86.6	3.1	97.9	2.2	89.1	84.9	5.2
Symbolic-only	100.0	0.0	92.9‡	93.5‡	23.9‡	95.3†	5.2†	75.2‡	76.1‡	32.2‡
FT-combin.§	–	–	–	–	–	–	–	87.0	82.8	11.7
Full pipeline	**100.0**	**0.0**	**96.9**	**95.8**	**1.2**	**98.4**	**1.1**	**91.1**	**90.1**	**5.3**

Table 11: Component ablation evaluated on the official test predictions: validity accuracy (%), premise retrieval F1 ($F1_p$), and TCE for each component on its own. †Symbolic-only on ST3 is computed over 190/192 parseable items; the 2 unparsed instances are non-English. ‡On ST2/ST4 the cascaded verifier (§1) invokes the parser only when FT confidence is below the bypass threshold or FT predicts VALID, so the parser produces a verdict on a subset of items only (100/192 for ST2, 109/192 for ST4); the reported numbers are computed over those subsets and therefore reflect parser behavior on the FT-uncertain portion of the test set rather than a hypothetical standalone run. §FT-combin.: FT model evaluates all $\binom{N}{2}$ premise pairs without the symbolic parser.

labeled as invalid. Major premise, minor premise, and conclusion are highlighted; filler premises are in black. †Valid only under existential import (Aristotelian logic).

1. **a660d443** (ST4/ru, Barbara AAA-1)
   Any snail moves with great speed. Every single sloth is a fast runner. It is true that all starfish are fast. Anything that is an animal is fast. Some clams are swift animals. Every single turtle is an animal. ∴ There are no turtles that are not fast.

2. **9d57ec74** (ST4/de, Celarent EAE-1)
   All zoos are empty of life. No mammal has a heart. Not a single creature that is a dog is an animal. Some bears are actually robots. Every single mammal is a dog. ∴ No mammal is an animal.

3. **827501ed** (ST4/es, Barbara AAA-1)
   Every single broccoli is a vegetable. It is true that all fruits are edible. Any potato is a vegetable. Some edible things are grains. There are a few roots that are edible. Every single vegetable is edible. Anything that is a carrot is a vegetable. ∴ There are no carrots that are not edible.

4. **475136be** (ST4/it, Barbara AAA-1)
   Every single elephant is capable of flight. There are no hippos that cannot fly. Any creature that is a dog can fly. Anything which is an animal is capable of flight. Every single insect is an animal. ∴ There are no insects that cannot fly.

5. **25d671a3** (ST4/ru, Baroco AOO-2)
   Every single person has three heads and five arms. It is the case that books eat the students who read them. No human has ever learned to read or write. Every person is a human. Some students are actually alien robots. At least one student is not a human. ∴ There is at least one student is not a person.

6. **4c62aa84** (ST4/sw, Baroco AOO-2)
   Many animals are kept by humans as companions. Some cats prefer to sleep for many hours a day. Every veterinarian treats various types of pets. Every single dog has fur. It is true that dogs are often called man's best friend. At least one cat does not have fur. ∴ At least one cat is not a dog.

7. **a3a4fcb5** (ST4/it, Celarent EAE-1)
   Not a single thing that is a vertebrate is a mammal. Some animals lay eggs in jelly. Every ecosystem has a food chain. It is true that skin can be permeable. Every single frog is a vertebrate. ∴ It must be the case that no frog is a mammal.

8. **37190acf** (ST2/en, Subaltern†)
   Every pea is a vegetable. Some beans are vegetables. There are many beets that are vegetables. Any radish is a vegetable. It is a fact that all turnips are vegetables. It is a fact that every potato is a vegetable. It is also true that all carrots are vegetables. ∴ A few vegetables are carrots.

9. **612f04e8** (ST2/en, Subaltern†)
   Some hinges are bones of a skeleton. All of the things that are doors are planets. There are no frames that are not made of flesh. It is true that all knobs are eyes of a beast. Any

ID	ST / Lang	GT	Form	Premises ⇒ Conclusion
Labeled invalid, instantiate valid forms (10)				
a660d443	ST4 / ru	inv	Barbara	All A are B; All C are A ⇒ All C are B
9d57ec74	ST4 / de	inv	Celarent	No A is B; All C are A ⇒ No C is B
827501ed	ST4 / es	inv	Barbara	All A are B; All B are C ⇒ All A are C
475136be	ST4 / it	inv	Barbara	All A are B; All C are A ⇒ All C are B
25d671a3	ST4 / ru	inv	Baroco	All A are B; Some C are not B ⇒ Some C are not A
4c62aa84	ST4 / sw	inv	Baroco	All A are B; Some C are not B ⇒ Some C are not A
a3a4fcb5	ST4 / it	inv	Celarent	No A is B; All C are A ⇒ No C is B
37190acf	ST2 / en	inv	Subaltern†	All A are B ⇒ Some B are A
612f04e8	ST2 / en	inv	Subaltern†	All A are B ⇒ Some A are B
0a7f2135	ST4 / ru	inv	Darapti†	All A are B; All A are C ⇒ Some C are B
Labeled valid, actually invalid (2)				
c6c1a6bc	ST4 / nl	val	Undist. mid.	All A are B; All C are B ⇏ All C are A
66efe831	ST4 / sw	val	4th-term	No A is B; Some B are C ⇏ Some A are not D

Table 12: The 12 annotation inconsistencies identified in the test set (§5.3). †Valid only under existential import (Aristotelian logic). The first 10 are false negatives (valid syllogisms mislabeled as invalid); the last 2 are false positives.

house that is built is actually a zoo. All of the things that are windows are planets. ∴ It can be deduced that there are windows that are planets.

10. **0a7f2135** (ST4/ru, Darapti†)
Some turtles are creatures capable of flight. Every single lizard is a mammal. It is the case that all frogs are reptiles. There are no reptiles that are cold-blooded. Anything that is a snake is a bird. It is a fact that all snakes are reptiles. ∴ It follows that a few reptiles are birds.

False positives. The 2 entries below show valid-labeled syllogisms that are actually invalid. Conclusions are highlighted; premises are in black to emphasize that no valid inference leads to the conclusion.

1. **c6c1a6bc** (ST4/nl, Undistributed Middle)
All humans are mammals with backbones. Any creature that is a dolphin is a mammal. Anything that has a backbone is a mammal. There are no dogs that do not have a backbone. It is true that some animals with backbones are reptiles. Every single fish has a backbone. Every single whale is a mammal. ∴ Every whale has a backbone.

2. **66efe831** (ST4/sw, Fourth Term)
It is a fact that animals adapt. Every zoo cares for wildlife. All farms have livestock. Most pets live indoors with humans. It is not true that any aquatic animals are cats. A portion of cats are domesticated animals. ∴ There are some aquatic animals that are not wild animals.

H Parser–FTM Synergy Details

This appendix expands the agreement matrix of Table 3 with per-language disagreement breakdowns and two case studies that illustrate how the two components correct each other. All counts are derived from the official test predictions, comparing the parser verdict and the FT verdict for each instance against the ground-truth label.

Per-language disagreement breakdown (ST3). Of the seven ST3 parser–FTM disagreements, six are FT-saves-Parser cases distributed across Russian, Swahili, French, Italian, Dutch, and Chinese (1 each); the single Parser-saves-FT case is Bengali. The three Both-wrong items are all in Russian (1) and Swahili (2). The disagreement pattern is therefore not concentrated in one language family; both Latin-script and non-Latin-script languages contribute.

Per-language disagreement breakdown (ST4). ST4's 26 parser–FTM disagreements split as 8 Parser-saves-FT (te, ru, fr, sw, te, fr, te, sw) and

18 FT-saves-Parser (most languages contribute 1–3 each, with Bengali, Italian, Dutch and Spanish at 2 each). The 9 Both-wrong cases concentrate in Russian (3), Italian (2), Spanish (1), German (1), Bengali (1), and Swahili (1), the same languages that dominate the annotation-inconsistency list of §G, supporting the reading that the Both-wrong cell is largely a label-noise artefact rather than a system failure.

Case study: Parser saves FT (ST2, id ba75f995). The 7-premise input concludes with a long-form "democracies are governing systems" chain. After hard stripping the relevant pair reduces to *All A are B; All C are A* $\Rightarrow$ *All C are B* (Barbara), which the symbolic parser confirms as VALID. The FTM, distracted by the dense surface form, returns INVALID with confidence 0.62, below the 0.85 bypass threshold, so the parser verdict is preserved and the prediction is correct. This is the canonical case for which the asymmetric-threshold design was engineered.

Case study: FT saves Parser (ST2, id 8dbac007). Premises [1] (*Anything that is a magazine is also something that can be eaten*) and [4] (*A few magazines are also books*) yield a valid Disamis: *All A are B; Some A are C* $\Rightarrow$ *Some C are B*. The combinatorial symbolic search finds no valid pair: the stripping output for premise [4] failed to preserve the existential quantifier in a parse-stable form, so no match was made in the 24-form lookup. The FTM correctly recognises the inference at confidence 0.989, the cascaded verifier accepts the FT verdict, and the final prediction is correct. This case illustrates the limitation that motivates the FT-primary design of the cascaded verifier: when stripping introduces surface-level noise, FT calibration on similar training distributions provides the recovery path the parser cannot.