YNJTC at SemEval-2026 Task 11: A Neuro-Symbolic Hybrid Pipeline for Content-Independent Syllogistic Reasoning

Junhao Fu* **Yun He** **Lina Zhao** **Weijuan Li**
Yunnan Communications Vocational and Technical College
Kunming, China
{fujunhao, heyun, zhaolina, weijuan.li.work}@ynjtc.edu.cn

Abstract

This paper presents a neuro-symbolic hybrid pipeline for SemEval-2026 Task 11 that addresses the challenge of disentangling content from formal reasoning in large language models (LLMs) by leveraging multilingual syllogistic reasoning. A key challenge in this task is the *content effect*, in which LLMs tend to conflate the plausibility of an argument's content with its logical validity. Previous approaches have relied predominantly on end-to-end neural models or direct LLM prompting, both of which remain susceptible to content biases acquired during pre-training. In this work, a three-layer hybrid pipeline is proposed that combines rule-based regular expression (regex) parsing, LLM-powered symbolic reasoning via lookup-table verification, and neural model ensembles as fallback classifiers. The core innovation lies in converting natural-language syllogisms into formal mood-figure representations via LLM parsing, then determining validity by consulting a fixed table of 24 classically valid syllogistic forms, thereby bypassing content-dependent reasoning entirely. The proposed system achieved a perfect Combined Score of 100.0 on Subtask 1 (English binary classification), with competitive results across all four subtasks. Comprehensive ablation experiments demonstrate that each pipeline layer contributes measurably to overall performance, and that the symbolic approach achieves near-zero content effect across multiple LLM parsers.

1 Introduction

SemEval-2026 Task 11 (Valentino et al., 2026) studies content-independent reasoning for syllogistic validity, including multilingual settings. A syllogism is valid when its conclusion follows from two premises due to form alone. In this task, the decision should not depend on whether the statements sound realistic.

A key difficulty is the *content effect*. Humans and LLMs often rate arguments as *more valid* when the content aligns with real-world knowledge, even when the form should matter (Dasgupta et al., 2022; Eisape et al., 2024). They can also be too forgiving when an invalid argument sounds plausible. This mixing of content and form makes logic judgments less reliable in applications that need strict validity checks.

Several lines of work aim to reduce content effects. One direction combines neural models with symbolic logic to make reasoning more faithful (Quan et al., 2024; Xu et al., 2024). Another uses prompting, such as chain-of-thought, to encourage structured intermediate steps (Lyu et al., 2023). There are also training-based approaches, including fine-tuning with demonstrations and activation steering, that target bias directly (Valentino et al., 2025). Even so, fully separating content from form remains difficult, especially across languages.

This paper presents a neuro-symbolic pipeline with a different strategy. Instead of forcing a model to *ignore* content, the input is converted into a formal structure, and a symbolic lookup table decides validity. The system has three layers. A regex rule module handles standard forms. An LLM-based parser extracts A/E/I/O types and term pairs, which are then mapped to mood and figure for table verification. If symbolic parsing fails, a neural ensemble is used as a fallback (DeBERTa-v3-large for English; XLM-RoBERTa-large for multilingual).

The main benefit comes from the lookup step. The 24 valid syllogistic forms depend on mood and figure, not on term meaning. The reported results show a Combined Score of 100.0 on Subtask 1 and scores of 54.43, 82.59, and 30.31 on Subtasks 2–4, respectively. Ablation results indicate that each layer contributes to overall performance. In addition, the content-effect measure-

* Corresponding author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 367–372

ments remain low for the symbolic part (TCE in the range 0.7–1.4% in later analysis).

The remainder of this paper is organized as follows. Section 2 describes the overall system architecture. Section 3 presents the experimental setup, and Section 4 reports results and analyses, including pipeline ablation, LLM parser comparisons, as well as content-effect analysis. Section 5 concludes the paper.

2 System Description

The proposed system employs a three-layer cascading pipeline architecture, as illustrated in Figure 1. Each layer attempts to determine the validity of a given syllogism; if a layer succeeds, its result is returned directly without invoking subsequent layers. This design prioritizes the most interpretable and content-independent methods, with neural models serving only as a safety net when symbolic reasoning fails. For Subtasks 2 and 4, which require identifying relevant premises from distractor sentences, an additional retrieval step is prepended to the pipeline, using LLM-based semantic analysis with keyword-matching fallback.

2.1 Layer 1: Regex-Based Rule System

The first layer uses deterministic regular expression matching to parse syllogisms into their formal components. Four proposition types are recognized: A (universal affirmative, e.g., **All X are Y**), E (universal negative, e.g., **No X are Y**), I (particular affirmative, e.g., **Some X are Y**), and O (particular negative, e.g., **Some X are not Y**). When all three propositions are successfully parsed, the system identifies the subject, predicate, and middle terms, computes the mood (a three-letter string of proposition types) and figure (1–4, determined by the position of the middle term), and consults the lookup table of 24 valid syllogistic forms.

This layer is restricted to Subtask 1, as it requires standard English syllogistic phrasing. Its primary advantage is perfect determinism: when parsing succeeds, the result is guaranteed to be content-independent. In practice, this layer covers approximately 37.7% of test samples and handles the most straightforwardly phrased syllogisms.

2.2 Layer 2: LLM-Powered Symbolic Parsing

Regex rules do not cover some syllogisms, including non-standard wording, complex noun phrases, and multilingual inputs. In these cases, the second layer uses an LLM as a structured extractor. DeepSeek-V3.2 (`deepseek-chat`) is prompted to return JSON with A/E/I/O types and subject–predicate pairs for the three propositions.

The JSON output is then passed to the same symbolic steps used in Layer 1. Mood is computed, and figure is determined via fuzzy term matching to handle simple form changes (e.g., singular vs. plural). The final label is obtained by table lookup over the 24 valid forms. In this layer, the LLM is used solely for parsing. The lookup table remains the only decision rule for covered samples.

This layer covers 91.2% of English samples and 71.4% of multilingual samples. On covered cases, symbolic accuracy is 96.0%. Coverage is lower in multilingual settings, consistent with the greater difficulty of structured parsing outside English.

2.3 Layer 3: Neural Model Ensemble

The third layer provides a neural fallback for samples where symbolic parsing fails (approximately 5–35% depending on subtask). DeBERTa-v3-large (He et al., 2023) with LoRA (Hu et al., 2022) is used for English subtasks, and XLM-RoBERTa-large (Conneau et al., 2020) with LoRA for multilingual subtasks. Five seeds are trained (13, 21, 42, 87, 111); seed 111 is dropped due to poor validation, yielding a four-seed ensemble with probability averaging. Training uses learning rate 2e-4, effective batch size 16, up to 15 epochs with early stopping (patience = 7), LoRA r=16, α=32. To mitigate content bias, R-Drop regularization (Liang et al., 2021), plausibility-aware supervised contrastive learning, adversarial debiasing with gradient reversal, and quadrant-aware data augmentation (3× base, 5× for hard cases) are applied. As a fallback layer, neural models process only symbolically unresolved samples, limiting their impact on overall bias.

3 Experimental Setup

Dataset. The task provides a training set of 960 English syllogisms, balanced across four quadrants defined by validity (valid/invalid) and plausibility (plausible/implausible): VP (240), VI (240), IP (234), and II (246). Test sets contain 191–192 samples per subtask. Subtasks 1 and 2 are in English; Subtasks 3 and 4 include multilingual syllogisms.

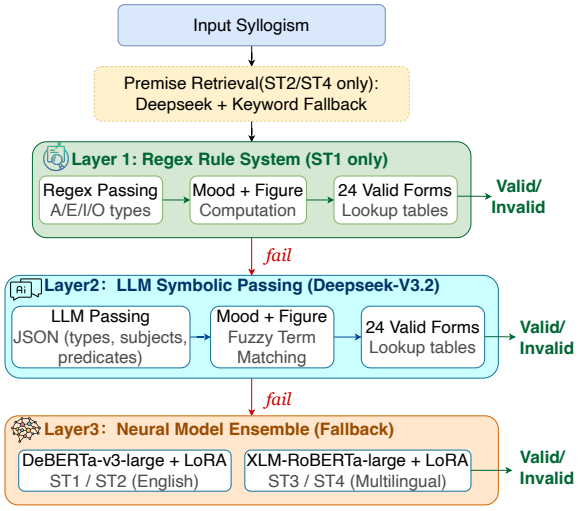

Figure 1: System architecture of the three-layer neuro-symbolic hybrid pipeline.

Evaluation Metrics. Performance is measured by the Combined Score, which rewards both high accuracy and low content bias:

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

$$\text{TCE} = |\text{Acc}_{\text{plaus}} - \text{Acc}_{\text{implaus}}|$$

where TCE (Total Content Effect) captures the gap between plausible and implausible accuracy.

Implementation Details. The LLM symbolic parsing layer uses DeepSeek-V3.2 via API with temperature = 0. Neural models are fine-tuned on a single NVIDIA 4090 GPU. The regex rule system uses a static lookup table of 24 valid syllogistic forms compiled from classical logic. All code is publicly available.[1]

4 Results and Analysis

4.1 Main Results

Table 1 presents the official evaluation results. The proposed system achieved a perfect Combined Score of 100.0 on Subtask 1, ranking joint first among all participants. On Subtask 2, a score of 54.43 was obtained (rank 4), where the additional challenge of identifying relevant premises from distractor sentences introduced greater variability. Subtask 3 yielded a score of 82.59 (rank 5), demonstrating that the symbolic parsing approach transfers effectively to multilingual settings. On Subtask 4 (multilingual retrieval), a score of 30.31 was achieved (rank 8), reflecting the combined difficulty of multilingual parsing and premise retrieval.

ST	Description	Rank	Score
1	English Classification	#11*	100.0
2	English Retrieval + Class.	#4	54.43
3	Multilingual Classification	#5	82.59
4	Multilingual Retrieval + Class.	#8	30.31

Table 1: Official evaluation results. *Tied first: 11 teams achieved 100.0 on ST1.

Configuration	Layers	Score	Δ
Full Pipeline	R + D + N	100.0	—
DeepSeek + Neural	D + N	57.38	−42.62
Neural Only	N	39.81	−60.19
DeepSeek Only	D	34.09	−65.91
Rules Only	R	13.70	−86.30

Table 2: Pipeline ablation on Subtask 1 (Combined Score). R = Rule system, D = DeepSeek symbolic, N = Neural ensemble. Uncovered samples default to Invalid.

4.2 Pipeline Layer Ablation

To quantify the contribution of each pipeline layer, ablation experiments were conducted by systematically removing individual components. Table 2 presents the results for Subtask 1, where all three layers are active, providing the most comprehensive ablation.

The results reveal a clear contribution from each layer. The rule system alone covers only 37.7% of samples (Score: 13.70), but when combined with the full pipeline, it contributes +42.62 points by handling the most straightforwardly phrased syllo-

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

ST	Full Pipeline	Neural Only	Δ
2	54.43	46.02	+8.41
3	82.59	37.32	+45.27
4	28.60	18.89	+9.71

Table 3: Symbolic layer contribution on Subtasks 2–4.

LLM	Par.	Cov.	Sym.	Acc	TCE	Score
DS-V3.2	99.9	91.2	96.0	91.4	0.8	57.03
Qwen3-Max	100	91.8	95.1	91.3	1.4	48.14
MiniMax	61.2	57.7	96.2	79.3	0.7	51.27
DS-V3.2-R	20.6	19.7	97.9	57.0	0.8	35.89

Table 4: Comparison of LLMs as symbolic parsers on training data (960 samples). Par. = Parse%, Cov. = Coverage%, Sym. = Symbolic accuracy. DS = DeepSeek, R = Reasoner. Uncovered samples default to Invalid.

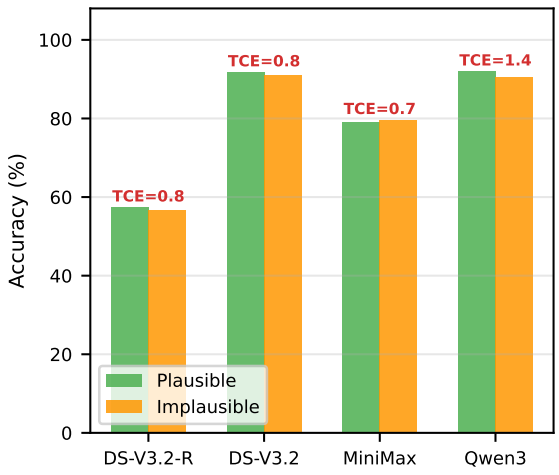

Figure 2: Content effect (plausible vs. implausible accuracy) across LLM parsers.

gisms with perfect accuracy. The DeepSeek symbolic layer adds +17.57 points by extending symbolic coverage to non-standard phrasings. The neural ensemble contributes a further +5.72 points as a safety net for samples that resist symbolic parsing.

The symbolic layer consistently improves performance across all subtasks (Table 3). The most substantial improvement is observed on Subtask 3 (+45.27), where multilingual XLM-RoBERTa alone achieves only 37.32, but the symbolic parsing layer raises the score to 82.59. This suggests that the symbolic approach is particularly valuable in multilingual settings, where neural models struggle with content-independent reasoning across diverse languages.

4.3 LLM Parser Comparison

Four LLMs were evaluated as parsers on the training set (960 samples) to test whether the parsing step generalizes across model families. The symbolic pipeline itself does not learn from these samples, since the lookup procedure is fixed. Table 4 reports the comparison.

A consistent pattern is observed. When structured parsing succeeds, symbolic accuracy stays above 95% across models. This suggests that the lookup step behaves similarly regardless of the upstream parser. The main difference comes from parsing success, which directly affects coverage. DeepSeek-V3.2 and Qwen3-Max achieve about 91% coverage, while MiniMax-M2.5 drops to 61.2% due to lower JSON parsing success rates.

DeepSeek-V3.2-Reasoner shows a different behavior. Its symbolic accuracy is high on covered cases (97.9%), but its parsing rate is low (20.6%). Many outputs include lengthy reasoning text rather than compact JSON, so the structured extraction step fails more often. This indicates that a reasoning-focused model may be less suitable as a strict information extractor than a general chat model, even if its reasoning ability is strong.

Across parsers, TCE stays low (0.7–1.4%). This is consistent with the design of the symbolic lookup step, where the label is determined by form rather than content.

4.4 Content Effect Analysis

This section analyzes content effects using plausible vs. implausible accuracy and a validity×plausibility quadrant breakdown. The same symbolic decision rule is used, while differences mainly reflect parser coverage and extraction stability. Figure 2 shows that all four parsers maintain low TCE (0.7–1.4%), confirming that the symbolic lookup step is inherently content-independent.

Table 5 reports accuracy for each parser under the four validity×plausibility quadrants. The Invalid+Plausible (IP) cases are important because they can look convincing even though they are invalid. Systems that rely on surface plausibility often fail in this quadrant.

Two patterns can be seen (Figure 3). For high-coverage parsers (DeepSeek-V3.2, Qwen3-Max), performance is relatively balanced across quadrants. The gap between quadrants stays within a small range (under 11 percentage points). This is

LLM	VP	VI	IP	II	Pl.	Im.	TCE↓
DS-V3.2	86.7	87.1	97.0	94.7	91.8	91.0	0.8
Qwen3	87.5	86.2	96.6	94.7	92.0	90.5	1.4
MiniMax	59.6	62.9	98.7	95.9	78.9	79.6	0.7
DS-V3.2-R	16.7	12.5	99.2	99.6	57.4	56.6	0.8

Table 5: Accuracy (%) by validity×plausibility quadrant. VP = Valid+Plausible, VI = Valid+Implausible, IP = Invalid+Plausible, II = Invalid+Implausible. Pl. = Plausible, Im. = Implausible.

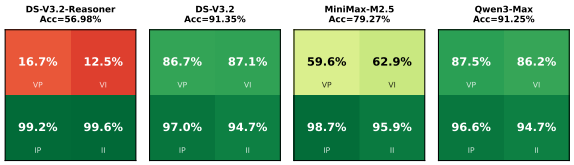

Figure 3: Accuracy by validity×plausibility quadrant for each LLM parser.

consistent with the intended behavior of the symbolic pipeline, where validity depends on the extracted form rather than plausibility.

Low-coverage parsers (DeepSeek-V3.2-Reasoner, MiniMax) show a different quadrant profile. IP and II are near-perfect (97–100%), while VP and VI drop substantially. This pattern is expected when uncovered samples default to Invalid: the default matches IP/II but harms VP/VI.

Figure 4 further suggests that most errors under DeepSeek-V3.2 come from figure calculation failures (83.5%), not from the lookup decision itself. This points to term matching and figure computation as the main practical bottleneck.

4.5 Case Study

Table 6 illustrates the pipeline on a valid-but-implausible syllogism, where the content may mislead neural models into predicting Invalid.

The regex layer parses all three propositions as type A, identifies the middle term *animals*, and computes mood AAA with figure 1. The lookup table confirms this as Barbara, one of the 24 valid forms. Because the decision depends entirely on form, the implausible content has no effect. A neural-only model, by contrast, is likely to reject the argument because the premises conflict with world knowledge.

5 Conclusion

This paper presented a neuro-symbolic pipeline that converts syllogisms into formal mood-figure representations and determines validity via symbolic table lookup, achieving a perfect score of 100.0 on Subtask 1. Ablation experiments confirmed that each pipeline layer contributes measurably, with the symbolic parsing layer providing the largest gains, particularly in multilingual settings (+45.27 on Subtask 3). Comparison across four LLM parsers showed that parsing success rate, rather than reasoning capability, is the primary bottleneck, and that all successfully parsed samples achieve near-zero content effect (TCE $\leq$ 1.4%). The case study further demonstrated that the symbolic approach correctly handles valid-but-implausible arguments that neural models tend to reject, highlighting the practical value of form-based reasoning. Future work will focus on improving multilingual parsing coverage and exploring lightweight specialized parsers to reduce dependence on LLM APIs.

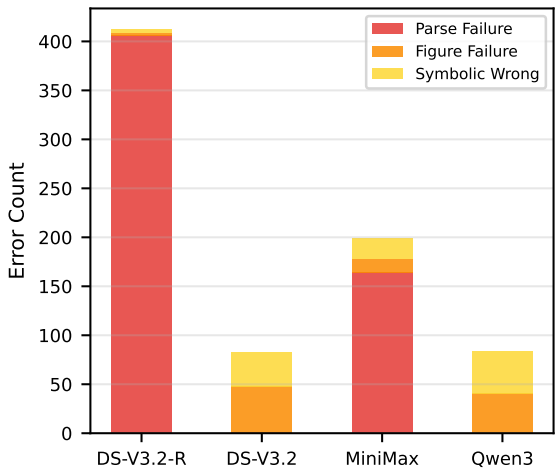

Figure 4: Error classification by type for different LLM parsers.

Input Syllogism
P1: All flowers are animals. P2: All animals are rocks. C: All flowers are rocks.
Pipeline Trace
Layer 1 (Regex): Parse → A, A, A. Terms: S = flowers, P = rocks, M = animals. Mood = AAA, Figure = 1 (M–P, S–M). Lookup: AAA-1 (Barbara) ∈ 24 valid forms → **Valid**
Neural baseline: Predicts Invalid (content is implausible).

Table 6: Pipeline trace on a valid-but-implausible syllogism. The symbolic lookup correctly identifies Barbara (AAA-1) regardless of content plausibility.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

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*.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.

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 (ICLR)*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

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.

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:2502.00407*.

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)*. 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)*, pages 13326–13365. Association for Computational Linguistics.