Team 0704mis at SemEval-2026 Task 11: Dual-View Consistency Testing for Content-Invariant Multilingual Syllogistic Reasoning

Ishita Gupta[1], **Dhruv Goyal**[2], **Dr. Jatin Bedi**[3]

Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology, Patiala, Punjab, India
isgupta0903@gmail.com, dhruv621999goyal@gmail.com, jatin.bedi@thapar.edu

Abstract

We describe our submissions to SemEval-2026 Task 11 Subtask 3, which is multilingual syllogistic validity classification over twelve typologically diverse languages. This task is challenging because of content effects in large language models, where semantic plausibility effects shift over formal logical validity effects in zero-shot cross-lingual transfer settings. We therefore pursue a neuro-symbolic architecture that ensures a hard boundary between the perception of natural language and the execution of logical reasoning. We implement a multilingual neural parser that categorically encodes simple textual syllogisms into logical forms, and a symbolic verifier that exactly implements the full set of 24 Aristotelian syllogistic forms, whose output determines validity.

To achieve robustness, we propose a dual-view consistency test, which uses two logical form abstraction views, native perceptual parsing and symbol-level content abstraction, and accepts only predictions where both views agree. This system achieves 95.83extensive ablation studies are that our failure of single-view parsing with native text on 72 cases (mostly on parsing failure for non-English languages) affected the coverage significantly. Our implementation is open-source available at: *https://github.com/09-ig/Semeval-task11-subtask3*

1 Introduction

Logic Meditation: The classical syllogism as a test suite for artificial reasoning systems. Formal Reasoning in Large Language Models: Evaluation in SemEval-2026 Task 11. Reasoning by logical inference requires determining whether conclusions follow from premises without attending to their content. Categorical syllogisms provide a controlled setting in which to study this ability: validity depends only on the formal structure. Recent work has shown that Large Language Models (LLMs) continue to show content effects, that is, judgments of syllogistic validity are influenced by semantic plausibility rather than logical form. Empirical work in cognitive science and NLP has shown that typical reasoners accept believable but invalid arguments and reject logically valid arguments with implausible conclusions. (Dasgupta et al., 2022) (Evans et al., 1983)

Task 11 of SemEval-2026 directly probes this phenomenon by requiring multilingual syllogistic reasoning. In particular, Subtask 3 requires systems to classify syllogistic validity in twelve typologically diverse languages. We approach this challenge with a neuro-symbolic architecture that separates perception from reasoning. A multilingual neural parser extracts categorical logical forms from natural language, and logical validity is then computed by a deterministic symbolic verifier that implements the complete set of 24 Aristotelian syllogistic forms. Since the verifier is provably complete, reasoning errors are eliminated, and remaining failures can be attributed to the instability of parsing.

Our novel contribution is dual-view consistency testing, that is, if the neural parser is asked to produce both the logical form for the native text and then again after a programmatic abstraction of all content symbols, then validity is accepted only when both views agree. This test allows us to determine whether the extracted logical form is indeed invariant to surface content, and avoids increasing the complexity of reasoning. Empirical analysis indicates that the dominant source of failures is indeed errors in multilingual parsing.

Contributions

This paper makes the following contributions: (1) A neuro-symbolic architecture that strictly separates neural parsing from symbolic verification, enabling precise error attribution. (2) Dual-view consistency checking as a content-sensitive pars-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1765–1771

ing detection mechanism with formal properties. (3) Extensive ablation across dual-view, single-view native, single-view masked, and local baseline configurations.

2 Background

2.1 Task Description

SemEval-2026 Task 11 Subtask 3 (Valentino et al., 2026) requires systems to perform binary classification of syllogistic validity across 12 languages. Each input instance consists of a natural language syllogism, two premises followed by a conclusion, in one of the target languages. The required output is a binary label: VALID if the conclusion follows necessarily from the premises by the rules of categorical syllogistic logic, INVALID otherwise. Crucially, validity depends *solely* on logical structure, independent of whether the premises or conclusion are true in the real world. Training data is provided exclusively in English, requiring systems to achieve zero-shot cross-lingual transfer to 11 additional languages. This typological diversity tests whether systems can extract logical structure from vastly different surface linguistic forms.

2.2 Formal Framework: Categorical Syllogisms

We follow the classical formalization of categorical logic as presented by Łukasiewicz (1957). Łukasiewicz's modern formal analysis of Aristotelian syllogistics is the groundwork of our symbolic verifier. A categorical syllogism consists of three propositions which include three terms: the major term(P), the minor term (S), and the middle term (M). Each proposition has 4 quantifier types: A: All S are P, E: No S are P, I: Some S are P, O: Some S are not P

The *logical form* of a syllogism is the pair $L = (m, f)$ consisting of its mood and figure. The *mood* of a syllogism is the ordered triple $m = (q_1, q_2, q_3) \in \{A, E, I, O\}^3$ specifying the quantifier types of the major premise, minor premise, and conclusion respectively. There are $4^3 = 64$ possible moods. The *figure* of a syllogism, $f \in \{1, 2, 3, 4\}$, specifies the arrangement of the middle term within the premises. Table 1 defines the four figures. Since there are 64 moods and 4 figures, there are $64 \times 4 = 256$ possible logical forms, of which exactly 24 are valid.

Theorem 1 (Completeness of Valid Forms). (Łukasiewicz, 1957) Let $V \subset \{A, E, I, O\}^3 \times \{1, 2, 3, 4\}$ denote the set of valid mood-figure combinations. Then $|V| = 24$, and a syllogism σ is logically valid if and only if its logical form $L(\sigma) \in V$. Table 2 enumerates all 24 valid forms.

Maj. Prem.	Min. Prem.	Middle Term Position
M – P	S – M	Subject in major, predicate in minor
P – M	S – M	Predicate in both premises
M – P	M – S	Subject in both premises
P – M	M – S	Predicate in major, subject in minor

Table 1: The four syllogistic figures, showing the position of major term (P), minor term (S), and middle term (M) in each premise.

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

Table 2: The 24 valid syllogistic forms with traditional names.

2.3 Related Work

Content Effects in Human Reasoning and LLM Large language models exhibit content effects in syllogistic reasoning. Cognitive psychology has established content effects in syllogistic reasoning. For instance, Evans et al. (1983) found that people tend to accept invalid but believable arguments and reject valid but unbelievable arguments. This is known as belief bias. Klauer et al. (2000) developed sophisticated dual-process models that distinguish between response bias (tendency to give certain responses) and reasoning bias (errors in the reasoning process itself), finding that belief affects both components. There is evidence that large language models are also subject to biases. Dasgupta et al. (2022) found LLMs systematically commit plausibility-driven reasoning errors, and Eisape et al. (2024) made detailed human–LLM comparisons. Ozeki et al. (2024) proposed controlled benchmarks to quantify content effects of neural systems.

Neuro-Symbolic Approaches to Reasoning. The integration of neural and symbolic methods for reasoning has received substantial recent attention. Pan et al. (2023) introduced Logic-LM, combining LLMs with symbolic solvers to improve reasoning accuracy. Lyu et al. (2023) proposed

faithful chain-of-thought reasoning that grounds intermediate steps in formal logic. Quan et al. (2024) developed methods for LLM-symbolic theorem prover collaboration, using the strengths of each component. Xu et al. (2024) advanced symbolic chain-of-thought approaches for faithful reasoning. Our work contributes to this literature by focusing specifically on the content effect problem and proposing dual-view consistency testing as a mechanism for detecting content-sensitive parsing.

3 System Overview

Our system implements a neuro-symbolic pipeline with four major components: (1) Multilingual Segmentation, (2) Neural Parsing, (3) Logical Form Extraction, and (4) Symbolic Verification.

3.1 Multilingual Segmentation

Syllogisms in natural language can be segmented into propositions in order to perform logical analysis. We use a lexicon based approach that applies in all 12 languages.

Segmentation Algorithm. The segmentation procedure operates as follows. (1) **Marker Search**: case-insensitively scan for conclusion markers in decreasing length order to prefer longer multi-word markers. (2) **Split**: content before the marker is premises; content after (excluding marker) is the conclusion. (3) **Premise Separation**: split premises on sentence-ending punctuation appropriate to the script (period for Latin/Cyrillic, danda for Bengali, etc.). (4) **Fallback**: if no marker is found, treat the last sentence as conclusion and all preceding sentences as premises.

3.2 Neural Parser

We use Claude Sonnet (`claude-sonnet-4-20250514`) via the Anthropic API to produce structured representations from natural language propositions with the neural parser. For consistency and reproducibility, the temperature $\tau = 0$.

Prompt Engineering. Prompt engineering is key to parsing correctness. Our prompt uses a range of techniques, following best-practice (Brown et al., 2020; Wei et al., 2022) using precise definitions (four types of quantifiers) and multilingual indicators (frequent expressions for each type). Further we use **Anti-Bias** examples to eliminate content effects. We have examples of implausible-but-valid syllogisms to indicate that plausibility is not relevant to the parsing pursuit:

```
(valid despite implausibility):
"All robots are clouds.
All cats are robots.
Therefore all cats are clouds."
```

We also explicitly teach the model to retrieve structure without validating it by **Explicit Task framing**.

```
IMPORTANT: Extract logical
  structure only.
Do NOT evaluate validity or
  real-world truth.
Plausibility is IRRELEVANT
  to this task.
```

textbfStructured Output Schema.: We specify a precise JSON schema for outputs including premises, quantifier and conclusion.

Schema Validation and Repair. We validate the parser output against the expected schema. We check: (1) structural validity, `premises` a list of two elements; (2) type validity, each proposition has a field `quantifier` with a value in $\{A, E, I, O\}$; (3) term validity, there are three distinct terms and the middle term occurs in both premises but not the conclusion. Upon validation failure we try one repair pass by sending the error message and a request to correct the output back to the model. This means a bounded computational cost, and we can recover from typical formatting errors.

3.3 Logical Form Extraction

From validated JSON output of the parser we apply a deterministic algorithm to extract the logical form $L = (\text{mood}, \text{figure})$.

The algorithm first identifies the three terms based on their roles in the conclusion (lines 1–7), then assigns premises to major/minor based on which contains the major term (lines 9–12), computes the figure from the middle term's position pattern (lines 14–19), and finally extracts the mood as the ordered triple of quantifier types (line 21).

3.4 Symbolic Verifier

The symbolic verifier checks the completeness theorem (Theorem 1) by hash table lookup. We store the 24 valid (mood, figure) pairs in a Python `frozenset` data structure. Verification is then a single membership test:

```
def verify(mood, figure):
    return (mood, figure) in VALID_FORMS
```

Algorithm 1: Logical Form Extraction

Input : Parsed propositions P_1, P_2, C from JSON
Output : Logical form (mood, figure) or FAIL

$S \leftarrow \text{subject}(C)$; $P \leftarrow \text{predicate}(C)$;
AllTerms $\leftarrow$ $\{\text{subj}(P_1), \text{pred}(P_1), \text{subj}(P_2), \text{pred}(P_2)\}$;
MidCands $\leftarrow$ AllTerms $\setminus \{S, P\}$;
if $|\textit{MidCands}| \neq 1$ **then return** FAIL;
$M \leftarrow$ the single element of MidCands;
if $P \in \{\textit{subj}(P_1), \textit{pred}(P_1)\}$ **then**
  MajorPrem $\leftarrow P_1$; MinorPrem $\leftarrow P_2$;
else
  MajorPrem $\leftarrow P_2$; MinorPrem $\leftarrow P_1$;
mjs $\leftarrow (\text{subj}(\text{MajorPrem}) = M)$;
mis $\leftarrow (\text{subj}(\text{MinorPrem}) = M)$;
if *mjs* $\wedge \neg$*mis* **then** fig $\leftarrow 1$;
else if $\neg$*mjs* $\wedge \neg$*mis* **then** fig $\leftarrow 2$;
else if *mjs* $\wedge$ *mis* **then** fig $\leftarrow 3$;
else fig $\leftarrow 4$;
mood $\leftarrow$ $(\text{quant}(\text{MajorPrem}), \text{quant}(\text{MinorPrem}), \text{quant}(C))$;
return (mood, fig);

This verifier has several important properties which includes **Completeness**, classifying every categorical syllogism correctly, no corner cases, no timeouts, no solver errors. **Efficiency** which refers to $O(1)$ time complexity via hash lookup, negligible compared to API latency. Then,**Content-Invariance** Works only with (mood, figure) pairs, unaware of natural language or content terms. Content-invariance is achieved by design. Finally, **Determinism**, when same input always produces the same output.

3.5 Dual-View Consistency Testing

Our official submission extends the base neuro-symbolic architecture with dual-view consistency testing, a mechanism for detecting content-sensitive parsing.

View A (Native Extraction). The original syllogism text is given to the parser together with a directive to extract the logical form and not to evaluate it on validity. This is the normal descriptive parsing procedure.

View B (Masked Extraction). The parser receives the same syllogism as in view A, but with different instructions:

"First, mentally replace all content nouns in the following syllogism with abstract symbols (X, Y, Z) while preserving the quantifier structure. Then extract the logical form from this abstracted representation."

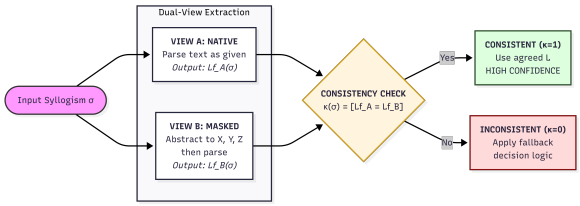

Figure 1: overview of Dual-View Consistency testing pipeline

Consistency Indicator. For syllogism σ, define:

$$\kappa(\sigma) = \mathbb{1}[L_A(\sigma) = L_B(\sigma) \wedge L_A(\sigma) \neq \bot] \quad (1)$$

where $\bot$ denotes parsing failure. That is, $\kappa(\sigma) = 1$ if both views succeed and produce identical logical forms. For **Decision Logic,** We handle different consistency scenarios: In **Case 1** ($\kappa = 1$, Consistent): Both succeed and agree; use the shared logical form. It is observational evidence of content-invariant parsing. In **Case 2** (Single View Available): One view succeeds; use it. Confidence is reduced as content invariance cannot be tested. In **Case 3** (Disagreement): Both views succeed but produce different LFs. Disagreement is interpreted as a manifestation of parsing instability rather than as evidence for the reliability of either view, and hence we rely on the masked view, which is more resistant to content effects, simply because it contains less content. Confidence is reduced because the invariance of the structural configuration is not secure. In **Case 4** (Both Fail): No valid parse. Fallback: direct validity query to LLM. Content effects may re-emerge, but is covered.

4 Experimental Setup

4.1 Data

We used the official SemEval-2026 Task 11 Subtask 3 dataset (**?**). The training set only includes English-language syllogisms, while the test set includes all 12 target languages with instances distributed evenly across validity and plausibility conditions.

4.2 Implementation Details

Neural Parser: We employ the Claude Sonnet (`claude-sonnet-4-20250514`) model available through the Anthropic API.[1] Parsing calls are made with temperature $\tau = 0$ to deterministically parse the input. We limit responses to a maximum of 1,000 tokens for parsing calls and 50 to-

[1] https://www.anthropic.com/api

kens for fallback validity checks. We employ retry logic with exponential backoff to handle rate limits. **Symbolic Verifier:** Implemented in Python 3.10 as a `frozenset` containing the 24 valid (mood, figure) tuples. Mood is a tuple of three single-character strings; figure an integer 1–4. Checking membership is $O(1)$. **Infrastructure:** All experiments were executed single-threaded to guarantee deterministic API responses.

5 Results

5.1 Official Competition Results

Table 3 presents our official submission results on the SemEval-2026 Task 11 Subtask 3 evaluation.

Metric	Value
Accuracy	**95.83%**
Total Content Effect (TCE)	5.21
Ranking Score (Accuracy / TCE)	**33.91**
Official Leaderboard Rank	**8th**

Table 3: Official results for our dual-view system on the Subtask 3 evaluation.

Our system attains a high accuracy (95.83%) with moderate content effect (5.21), and produces a competitive gain-adjusted ranking score (33.91). TCE refers to the extent to which predictions are driven into the wrong direction by plausibility as opposed to logical form. The difference in accuracy between plausible but invalid syllogisms and implausible but valid syllogisms. The smaller the TCE, the greater the content invariance.

5.2 Ablation Study Results

Table 4 compares different system configurations to isolate component contributions.

Configuration	Accuracy	TCE	Score
Dual-View (Official)	**95.83%**	5.21	**33.91**
Single-View Native	92.08%	4.89	29.11
Local Baseline	68.33%	24.17	2.83

Table 4: Ablation study results comparing different system configurations.

72 parsing failures occurred with single-view native parsing on the test set. These failures were concentrated in non-English languages, especially those with complex morphology or non-Latin scripts. Rule-based pattern matching without neural (syntactic) parsing is significantly worse (68.33% accuracy, 156 failures) guaranteeing that sophisticated language understanding is needed for processing multilingual syllogisms.

After the official evaluation period, we conducted additional experiments to better understand our system's behavior. Table 5 presents post-competition results.

Configuration	Acc.	TCE	Score
Single-View Masked	**97.50%**	**1.71**	**57.00**

Table 5: Post-competition results for single-view masked configuration.

6 Discussion

Our results support the conjecture that abstraction and consistency constraints act on neuro-symbolic reasoning in complementary ways. Masked-only achieves higher accuracy and lower content effect due to the regularizing effect of symbol abstraction, while dual-view consistency primarily serves as a diagnostic mechanism for detecting parsing instability rather than maximizing raw accuracy. Symbol abstraction benefits as a structural regularizer, while dual-view consistency may facilitate a general diagnostic approach to detecting content-sensitive parsing instability and providing principled failure attribution throughout the pipeline. Three caveats necessitate careful interpretation: (1) **Scope**: we only treat classical two-premise categorical syllogisms; (2) **Post-competition evaluation**: our masked-only result (57.00) was evaluated on the same test set after official evaluation; (3) **API dependence**: using the commercial LLM limits the reproducibility and the drive to an open-source substitute.

7 Conclusion

We describe our submission to SemEval-2026 Task 11 Subtask 3: system for multilingual classification of logical validity in 12 languages with dual view consistency testing and formal characterization. Our system enforces separation of neural parsing and symbolic checking, which permits content effects only via parsing (which is observable). **Future Work** directions include: (1) improved masked parsing prompts to further reduce content effects; (2) robust consistency checks tolerant of minor formatting variations; and (3) extension beyond categorical syllogisms to propositional and first-order logic.

References

Tom Brown, Ben Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, 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 Thomas L. Griffiths. 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*, pages 3274–3289, Mexico City, Mexico.

Jonathan St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. volume 11, pages 295–306.

Karl Christoph Klauer, Jochen Musch, and Berthold Naumer. 2000. On belief bias in syllogistic reasoning. *Psychological Review*, 107(4):852–884.

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*, pages 305–329, Nusa Dua, Bali.

Kentaro Ozeki, Risako Ando, Terufumi Morishita, Jun Suzuki, and Masaaki Nagata. 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 8734–8756, Bangkok, Thailand.

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

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through LLM-symbolic theorem prover collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12345–12360, Miami, Florida.

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)*.

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.

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 Complete System Hyperparameters

Parameter	Value
Model	`claude-sonnet-4-20250514`
Temperature	0
Max Tokens (Parsing)	1000
Max Tokens (Fallback)	50
API Rate Limit Delay	50 ms
Retry Attempts	3
Schema Repair Attempts	1
Consistency Threshold	Exact match

Table 6: Complete system hyperparameters.

B Prompt Templates

B.1 Native Parsing Prompt (View A)

```
You are a logical form extractor for
categorical syllogisms.

QUANTIFIER TYPES:
A (Universal Affirmative): "All S
are P"
E (Universal Negative): "No S are P"
I (Particular Affirmative):
"Some S are P"
O (Particular Negative): "Some S
are not P"

IMPORTANT:
Extract logical structure only.
Do NOT evaluate validity.
Plausibility is IRRELEVANT to
this task.

ANTI-BIAS EXAMPLE (valid despite
implausibility):
"All robots are clouds.
All cats are robots.
 Therefore all cats are clouds."
 -> AAA-1

OUTPUT ONLY JSON:
{
  "premises": [
    {"quantifier": "A/E/I/O",
```

```
    "subject":
    "...", "predicate": "..."},
    {"quantifier": "A/E/I/O",
    "subject":
    "...", "predicate": "..."}
  ],
  "conclusion":
    {"quantifier": "A/E/I/O",
    "subject":
    "...", "predicate": "..."}
}

SYLLOGISM: {input}
```

B.2 Masked Parsing Prompt (View B)

```
[Same as View A, with the following additional instruction:]

TASK:
First, mentally replace all content nouns in the syllogism
with abstract symbols (X, Y, Z) while preserving quantifier
structure. Then extract the logical form from this abstracted
representation.
```