0704mis at SemEval-2026 Task 11: Single-Call Joint Abstraction for Robust Neuro-Symbolic Retrieval

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, dhruv62199goyal@gmail.com, jatin.bedi@thapar.edu

Abstract

We present a neuro-symbolic pipeline achieving ranking scores in SemEval-2026 Task 11 Subtasks 2 (English) and 4 (Multilingual) for joint validity classification and premise retrieval of syllogistic reasoning with irrelevant distractor premises. Our key design principle is single-call joint abstraction: a single LLM call jointly parses all premises and the conclusion, extracting symbols (X, Y, Z, etc) to abstract content terms and categorical logical forms (A/E/I/O) to abstract logical operators/relations. Processing joint necessitates that the same entities be abstracted to the same symbols across all statements, which is essential for the ability to detect valid syllogistic structure. An exhaustive pair search then checks all $O(n^2)$ premise pairs against the 24 valid Aristotelian forms via $O(1)$

Extensive ablation studies comparing our submitted system to variants that are theoretically more sophisticated (Form First Retrieval (FFR) with structural confidence metrics and Symbolic Entailment Search (SES) with formal entailment kernels) indicate that when parsing is the bottleneck, simplicity and robustness outweigh theoretical sophistication. https://github.com/09-ig/Semeval-task11-subtask2-4

1 Introduction

1.1 The Premise Retrieval Challenge

Subtasks 2 and 4 in SemEval-2026 Task 11 (Valentino et al., 2026) present a novel challenge connecting information retrieval, formal logic, and natural language understanding. In standard information retrieval, relevance is about semantic similarity, topical coherence, or keyword overlap. This premise retrieval task concerns *logical relevance*—structure and formal relationships, not content.

Formally: given a syllogism $\sigma = (P_0, P_1, \ldots, P_{n-1}, C)$ composed of n premises (including logically irrelevant distractors) and a conclusion C, the system must output (1) a validity judgment specifying whether any two-premise subset entails the conclusion, and (2) if valid, the indices of the necessary and sufficient premises. This captures logical reasoning with noise, a realistic setting where not all available information is relevant. Large Language Models exhibit systematic content effects in logical reasoning: like humans, they conflate semantic plausibility with logical validity, reflecting well-known cognitive biases. This was first documented in human reasoning by Evans et al. (1983) and in LLMs by Dasgupta et al. (2022). LLMs tend to over-endorse arguments with believable conclusions, regardless of structural validity.

1.2 Neural-Symbolic Interface Stability

For categorical syllogisms, every valid argument has exactly two premises. Retrieval is therefore the search for a 2-element minimal entailing subset (MES) within the set of all premise pairs. With accurate extraction of logical forms, validity is deterministic and exact. The challenge is at the neural–symbolic interface. Validity depends on detecting a shared middle term across premises. A consistent symbol across sentences is necessary to ensure correct structural detection

1.3 Our Approach: Robust Neuro-Symbolic Retrieval

We present a three-step neuro-symbolic pipeline:

Stage 1 : Multilingual Segmentation: Premises and conclusion are identified using rule-based markers and language specific punctuation (in subtask 4). **Stage 2 : Single-Call Joint Abstraction:** A single call to an LLM (Claude Sonnet, `claude-sonnet-4-20250514`) jointly processes all statements, outputting: (a) the conclusion sentence index; (b) for each sentence, the quantifier type (A/E/I/O) and abstract terms (X, Y, Z, ...). The single-call aspect is critical: if we

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1944–1949

parsed statements separately, the same entity could be assigned different symbols, breaking syllogistic structure detection. **Stage 3 : Exhaustive Pair Search:** Check all $\binom{n}{2}$ premise pairs against the 24 valid Aristotelian syllogistic forms. For each pair, mood and figure are computed and validated using $O(1)$ symbolic lookup. The search outputs the best valid pair, with tie-breaking by conclusion term coverage and index order.

Our contributions include a novel representation of the syllogistic premise retrieval problem as a search for minimal entailing subsets. A single call mechanism to enforce consistency of symbols across premises. A combinatorial search procedure with deterministic symbolic verification. Ablation studies with two formal variants– Form first retrieval and Symbolic Entailment Search.

2 Background and Task Formulation

2.1 SemEval-2026 Task 11

SemEval-2026 Task 11 (Valentino et al., 2026) assesses the degree to which language models can perform formal logical reasoning isolated from content effects. The task is motivated by the observation that despite impressive performance on reasoning benchmarks, LLMs often rely on pattern matching and semantic associations rather than genuine logical inference. We focus on the subtask 2 and 4. Their formal specification:

Input: A syllogism $\sigma = (P_0, P_1, \ldots, P_{n-1}, C)$ where $n \in [3, 7]$ includes logically irrelevant distractor premises.

Output: is given by a tuple (validity, relevant_premises) where validity is Boolean, describing whether any two-premise subset entails C, and relevant_premises is the list of the two necessary and sufficient premise indices if valid, or empty if invalid.

Constraints: If valid, return only the minimal 2-premise subset. Selecting only 1 premise yields zero score. If invalid, relevant_premises must be empty.

2.2 Categorical Syllogisms

A categorical syllogism is a deductive argument comprising three categorical propositions using three terms. First studied by Aristotle in *Prior Analytics*, it remains a central topic in logic for over two millennia. The modern formalization follows Łukasiewicz (1957). A categorical proposition has the form $\varphi = (q, S, P)$ where $q \in \{A, E, I, O\}$ is the quantifier type, S is the subject term, and P is the predicate term. A categorical syllogism consists of three categorical propositions: two premises and one conclusion, involving a major term, a minor term and a middle term. The mood 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 logical form is the pair $L = (m, f)$ consisting of mood and figure. Since there are 64 moods and 4 figures, there are $64 \times 4 = 256$ possible logical forms.

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 1: The 24 valid syllogistic forms with traditional names.

2.3 Related Work

Neuro-Symbolic Approaches. Pan et al. (2023) introduced Logic-LM, combining LLMs with symbolic solvers. Lyu et al. (2023) proposed faithful chain-of-thought grounding intermediate steps in formal logic. Our work differs in focusing on minimal entailing subset identification under distractors. The symbolic component is exact and exhaustive, performance is therefore validated primarily by the stability of logical-form extraction across multiple premises.

Premise Selection. Premise selection has been studied extensively in automated theorem proving, where the goal is to identify relevant axioms from large corpora. Approaches include, ranging in complexity, from kernel-based approaches to deep neural ranking models. However, these settings rely on semantic similarity and are based on large knowledge bases. Alama et al. (2014) studied premise selection for theorem proving using corpus analysis and kernel methods. Irving et al. (2016) applied deep sequence models for premise selection in mathematical reasoning. Our work differs by addressing premise selection in natural language syllogistic reasoning with explicit distractor premises.

3 System Description

3.1 Architectural Overview

The complete pipeline: (1) **Segmenter** detects conclusion markers and splits into $[P_0, P_1, \ldots, P_{n-1}, C]$; (2) **Single-call joint abstraction** with an LLM to obtain categorical forms and a globally consistent symbol mapping; (3) **Symbolic Validator** checks all $\binom{n}{2}$ premise pairs against the 24 valid forms via $O(1)$ hash lookup, outputting the best valid pair if any exist.

3.2 Stage 1: Multilingual Segmentation

Syllogisms in natural language must be segmented into propositions for logical analysis. We use a lexicon-based approach with language-specific conclusion markers and punctuation.

Segmentation Algorithm. The procedure operates as follows: (1) **Marker Search**: scan case-insensitively for conclusion markers in longest-match-first order to prefer 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). (4) **Fallback**: if fewer than 3 segments result or no marker is found, fall back to newline/semicolon splitting.

3.3 Stage 2: Single-Call Joint Abstraction

We use Claude Sonnet (`claude-sonnet-4-20250514`) via the Anthropic API to jointly process all statements with temperature $\tau = 0$ for determinism.

The Case for Single-Call Processing. Single call joint abstraction is critical for symbol consistency. If we parsed statements separately, the same entity could be assigned different symbols (e.g., "vehicles" could be X in one statement, Y in another), breaking syllogistic structure detection. Joint processing allows the model to maintain consistent abstractions throughout the argument.

For example: "All bicycles are vehicles. All vehicles have wheels. Therefore, all bicycles have wheels." Separate parsing might yield:
Premise 1: All X are Y (bicycles - X, vehicles - Y)
Premise 2: All Z are W (vehicles - Z, wheels - W)
Conclusion: All X are W (bicycles - X, wheels - W)

The middle term "vehicles" is inconsistently mapped (Y vs. Z), making structure detection impossible. Joint parsing ensures: vehicles - Y throughout, yielding the correct mood AAA and figure 1.

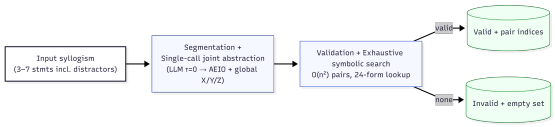

Figure 1: Single-call Joint Abstraction pipeline

Prompt Engineering. Our prompt employs several techniques following best practice (Brown et al., 2020; Wei et al., 2022) including **Precise Definitions** where we define the four quantifier types with canonical forms and semantic meanings. Then, we instruct the model to replace the content nouns with abstract symbols while maintaining the quantifier structure using **Joint Abstraction Instruction.**:

```
TASK: Process all statements jointly.
Replace all content nouns with abstract
symbols (X, Y, Z, ...) consistently in
ALL statements. Use the SAME symbol for
the SAME entity everywhere it appears.
```

Further, we provide **Anti-Bias Examples** with implausible content to emphasize structure over plausibility. Finally, We specify a precise JSON schema with conclusion index and structured logical forms. This design isolates structure from content and ensures compatibility with symbolic validation.

Output Processing and Validation. We validate parser output against the expected schema, checking: (1) **structural validity**: exactly one conclusion index when $n \geq 3$ statements; (2) **type validity**: each statement has a quantifier in $\{A, E, I, O\}$ and distinct subject/predicate; (3) **symbol validity**: only symbols X, Y, Z, etc are used (no residual content terms). If a validation fails we perform one repair pass by returning our error message to the model. This gives us bounded computational cost while maintaining the ability to recover from common formatting errors.

3.4 Stage 3: Exhaustive Pair Search

The final stage checks all $\binom{n}{2}$ premise pairs against the 24 valid Aristotelian syllogistic forms.

The Symbolic Validator. The validator stores the 24 valid (mood, figure) pairs in a Python frozenset:

```
def verify(premi1, premi2, conc):
    # Extract mood and figure
    mood =
```

```
    (premi1.q, premi2.q, conc.q)
    figure =
    compute_figure(premi1, premi2, conc)
    return (mood, figure) in VALID_FORMS
```

This verifier makes sure:
Completeness: It correctly classifies all categorical syllogisms. There should be no corner cases, no solver errors, and no timeouts. **Efficiency:** $O(1)$ time via hash lookup, negligible compared to the latency caused by API. **Content-Invariance:** Operates only on (mood, figure) pairs; unaware of natural language or content terms. **Determinism:** Same input always produces the same output

Exhaustive Search Algorithm. Algorithm 1 gives the complete procedure for exhaustive search.

Algorithm 1: Exhaustive Premise Pair Search

Input : Parsed statements $\{S_0, \ldots, S_{n-1}\}$ with conclusion index c
Output : (validity, relevant_premises)

$C \leftarrow S_c$; $\Pi \leftarrow \{S_i : i \neq c\}$; $n_p \leftarrow |\Pi|$;
valid_pairs $\leftarrow \{\}$;
for $i \leftarrow 0$ **to** $n_p - 1$ **do**
  for $j \leftarrow i + 1$ **to** $n_p - 1$ **do**
    $P_1 \leftarrow \Pi[i]$; $P_2 \leftarrow \Pi[j]$;
    $M \leftarrow$ identify_middle_term(P_1, P_2, C);
    if $M = \perp$ **then continue**;
    fig $\leftarrow$ compute_figure(P_1, P_2, M);
    mood $\leftarrow$ (quant(P_1), quant(P_2), quant(C));
    if *(mood, fig)* $\in V$ **then**
      valid_pairs.add$((i, j))$;
if *valid_pairs* $= \emptyset$ **then**
  return (False, []);
else
  best $\leftarrow$ select_best(valid_pairs, C);
  return (True, best);

The identify_middle_term function picks candidate terms that occur in both premises and not in the conclusion. For each candidate middle term M, the compute_figure decides the possible syllogistic figure (1–4) on the syntactic position that M occurs in the two premises. In figure 1, middle term M is subject of the major premise and predicate of the minor premise (M-P, S-M). In figure 2, M is a predicate of both premises (P-M, S-M). In figure 3, M is a subject of both premises (M-P, M-S). In figure 4, M is a predicate of the major premise and a subject of the minor premise (P-M, M-S). For each candidate configuration the corresponding mood is then computed from the quantifiers of both premises and the conclusion. A pair of premises is accepted if a candidate middle term yields a valid (mood, figure) combination among the 24 admissible forms.

Tie-Breaking Logic. When more than one valid premise pair is found, we choose one pair. Firstly, we compute a *coverage score* where we count how many of the conclusion's terms (subject and predicate) appear in the pair: 2 if both a subject and predicate are present, 1 if a single one is present, and 0 otherwise. We prefer higher coverage pairs. If multiple pairs achieve the same coverage score, we break ties by lexicographic index order, preferring lower indices (e.g., $(0, 1) < (0, 2) < (1, 2)$). This ensures consistent outputs across runs.

4 Ablation Studies

We created two variants for ablation studies with increased theoretical sophistication to test our robust principle.

4.1 Form First Retrieval (FFR)

Form First retrieval method introduces a confidence metric $c(\sigma)$ quantifying certainty based on the number of entailing pairs found:

Definition of Structural Confidence:

$$c(\sigma) = \begin{cases} 0 & \text{if } |R^*(\sigma)| = 0 \\ 1 - \frac{1}{|R^*(\sigma)|+1} & \text{if } |R^*(\sigma)| > 0 \end{cases} \quad (1)$$

where $R^*(\sigma)$ is the set of all valid premise pairs. The intuition is that confidence increases with multiple valid pairs, whereas one pair reflects intermediate confidence. The metric ranges from 0 (no valid pairs) to asymptoting to 1 (many valid pairs).

FFR also tracks symbolic coverage $\bar{\kappa}_\ell$ for each language ℓ, measuring the fraction of instances resolved symbolically (without fallback). It formalizes validation as an *entailment kernel*:

$$K(\varphi_1, \varphi_2, \varphi_C) = \mathbb{1}\Big[\big(\text{mood}(\varphi_1, \varphi_2, \varphi_C), \text{figure}(\varphi_1, \varphi_2, \varphi_C)\big) \in V\Big] \quad (2)$$

FFR triggers fallback based on confidence thresholds: if $c(\sigma) < \tau_1$, invoke additional parsing with modified prompt; if $\bar{\kappa}_\ell < \tau_2$, use ensemble parsing (multiple prompts, majority vote).

4.2 Symbolic Entailment Search (SES)

SES formalizes retrieval as finding a minimal entailing set:

Definition (Minimal Entailing Set). Given premises $\Pi = \{P_0, \ldots, P_{n-1}\}$ and conclusion C, the Minimal Entailing Set is:

$$\text{MES}(\Pi, C) =_{R \subseteq \Pi} |R| \text{ such that } R \models C \quad (3)$$

Theorem: For categorical syllogisms, $|\text{MES}| = 2$ always.

SES includes a complexity analysis proving that any algorithm for MES search must examine $\Theta(n^2)$ pairs in the worst case (an adversary can construct inputs where the valid pair is the last examined under any ordering), showing our search is asymptotically optimal. SES also provides a content-invariance theorem: Let π be the parsing function. If $\pi(\sigma_1) = \pi(\sigma_2)$ (identical logical forms), then $\text{SES}(\sigma_1) = \text{SES}(\sigma_2)$ (identical outputs). This formalizes the intuition that content effects can only enter through parsing, not search.

4.3 Experimental Comparison

We evaluated all three systems (RNSR, FFR, SES) on both Subtask 2 (English) and Subtask 4 (Multilingual) under identical conditions:

System	ST2	ST4
RNSR (Submitted)	**28.08**	**26.14**
Form First Retrieval (FFR)	27.26	19.30
Symbolic Entailment Search (SES)	20.85	17.86

Table 2: Ablation comparison showing RNSR outperforming both sophisticated variants. ST2 = Subtask 2 (English), ST4 = Subtask 4 (Multilingual).

5 Experimental Results and Analysis

5.1 Official Competition Results

Subtask	Accuracy	TCE	Score	Rank
2 (English)	91.58%	7.33	**28.08**	**11th**
4 (Multilingual)	76.56%	5.42	**26.14**	**10th**

Table 3: Official competition results for RNSR.

We achieve a competitive classification accuracy on both tasks (91.58% English, 76.56% multilingual), showing our neuro-symbolic pipeline effectively distinguishes valid from invalid syllogisms even with distractor premises. TCE scores are quite low (7.33 English, 5.42 multilingual), indicating strong content invariance. Achieving 11th place (English) and 10th place (Multilingual) in a competitive field demonstrates that our simple, robust approach is competitive with more complex systems while maintaining interpretability and reproducibility. The performance gap in the English and Multilingual settings is not caused by the limitations of symbolic reasoning but increased parsing complexity. Multilingual input induces variations in morphology, punctuation, and sentence structure, which lead to higher chances for segmentation and quantifier extraction errors. This is in line with our findings that errors arise at the neural–symbolic interface and not at the symbolic validation.

5.2 Error Analysis

A quick survey of unlabeled cases suggests that three problems account for most of the errors: misidentification of the conclusion sentence during the segmentation step, mislabelling of numerals in linguistically complex sentences and, rarely, inconsistency of symbols across the premises. The errors were particularly pronounced in the multilingual material, especially when the sentences had flexible word order or were written in non-Latin scripts.

6 Discussion

Scope Limitations. We restrict our focus to two-premise categorical syllogisms to allow for controlled evaluation of the interplay between logical-form extraction and symbolic inference. Our current system and pairwise search strategy can be applied to more complex logs, such as polysyllogisms and/or hypothetical syllogisms. We selected Claude Sonnet, as it was easier to coax structured output; however, the method is model-agnostic, and other language models with similar behavior can be plugged into the system, with the main difference being parsing power (as opposed to symbolic reasoning capability). **Future Directions.** Future work involves building dedicated logical-form parsers and designing for more robustness to non-classical quantifiers and multilinguality.

7 Conclusion

Our robust neuro-symbolic system for syllogistic premise retrieval achieved 11th place (Subtask 2: English, score 28.08) and 10th place (Subtask 4: Multilingual, score 26.14) in SemEval-2026 Task 11. Our approach (RSNR) applies single-call joint abstraction followed by exhaustive pair search against the 24 valid Aristotelian syllogistic forms.

Key Takeaways: Joint parsing in a single call ensures symbol consistency, which is critical for detecting syllogistic structure. Processing statements together allows the model to maintain consistent abstractions.Exhaustive search at $O(n^2)$ with fast $O(1)$ symbolic validation is tractable and effective.

References

Jesse Alama, Tom Heskes, Daniel Kühlwein, Evgeni Tsivtsivadze, and Josef Urban. 2014. Premise selection for mathematics by corpus analysis and kernel methods. *Journal of Automated Reasoning*, 52(2):191–213.

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 CY 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*.

Jonathan St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306.

Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Eén, François Chollet, and Josef Urban. 2016. DeepMath—deep sequence models for premise selection. In *Advances in Neural Information Processing Systems*, pages 2235–2243.

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.

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.

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. *Advances in Neural Information Processing Systems*, 35:24824–24837.