ITLC at SemEval-2026 Task 11: Normalization and Deterministic Parsing for Formal Reasoning in LLMs

Wicaksono Leksono Muhamad[*1,2], **Joanito Agili Lopo**[*1,2], **Tack Hwa Wong**[*1,3],
Muhammad Ravi Shulthan Habibi[1,4], **Samuel Cahyawijaya**[*1,5]
[1]SEACrowd [2]Mantera Studio [3]Universiti Teknologi PETRONAS
[4]Universitas Indonesia [5]Cohere
`{wcksnlxn,amalopo99,tackhwawong00}@gmail.com`
`muhammadravi251001@gmail.com`, `samuelcahyawijaya@cohere.com`
Code: `https://github.com/SEACrowd/ITLC_semeval2026_shared_task_11`

Abstract

Large language models suffer from content effects in reasoning tasks, particularly in multilingual contexts. We introduce a novel method that reduces these biases through explicit structural abstraction that transforms syllogisms into canonical logical representations and applies deterministic parsing to determine validity. Evaluated on the SemEval-2026 Task 11 multilingual benchmark, our approach achieves top-6 rankings across all subtasks while substantially reducing content effects and offering a competitive alternative to complex fine-tuning or activation-level interventions.

1 Introduction

The scope to which large language models (LLMs) can perform content-independent reasoning remains a central question in reasoning tasks. Prior work has shown that LLMs exhibit strong *content effects* in real-world knowledge and belief during pre-training (Dasgupta et al., 2024; Bertolazzi et al., 2024). These findings raise concerns about robustness, bias, and reliability in LLM applications.

Recent works have explored different mitigation methods to this problem. For instance, Kim et al. (2025) show that LLMs develop specific inference mechanisms in their internal architecture. Similarly, Valentino et al. (2025) introduce kNN-based conditional steering in the architecture to reduce content effect. Meanwhile, Neuro-symbolic and quasi-symbolic approaches have also been explored to improve faithfulness and logical consistency (Ranaldi et al., 2025; Quan et al., 2024; Xu et al., 2024; Lyu et al., 2023). Despite these advances, there is still no simple and effective solution for disentangling content from formal reasoning, particularly in multilingual settings.

In this work, we introduce a novel unbiased method for syllogistic reasoning that reduces content effects through explicit structural abstraction. Our approach transforms each argument into a canonical syllogistic representation that preserves only its logical structure, followed by deterministic structural parsing to determine validity. This simple strategy substantially reduces content effects while achieving strong validity accuracy.

We evaluate our method on the SemEval-2026 Task 11 (Valentino et al., 2026), a multilingual benchmark for syllogistic reasoning that explicitly measures both validity accuracy and the magnitude of content effects. Our method ranks in the top-5 across 3 subtasks among all participants and 6th position in subtask 2 (See Appendix A for detail leaderboard comparison). These results demonstrate that our structural abstraction approach remains a competitive and interpretable alternative to heavy fine-tuning (Ranaldi et al., 2025; Bertolazzi et al., 2024) or latent-level interventions (Valentino et al., 2025; Lopo et al., 2025) for mitigating reasoning biases in both English and multilingual settings.

2 Background

Categorical Syllogisms Categorical syllogisms are a compact form of deductive reasoning consisting of two premises and a conclusion (Prior, 1962; Ramsey, 2009; Priest, 2008). Their validity is entirely determined by structural configuration, making them a natural benchmark for evaluating whether models follow logical form rather than surface cues (Wu et al., 2023; Ozeki et al., 2024). In practice, the core challenge lies in mapping natural language text onto the intended quantifiers, negations, and term relations, a process that is brittle under paraphrase and compounds across languages (Zong and Lin, 2024; Cui et al., 2022).

Logical Structure and Terminology A categorical syllogism uses three terms. The subject of the conclusion is the minor term (S), the predicate of the conclusion is the major term (P), and the term

---

[*] Equal contributions.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1730–1747
July 3-4, 2026

that appears in both premises but not in the conclusion is the middle term (M). For a valid syllogism, the conclusion is always a claim about the relation between S and P (Eisape et al., 2024). In other words, M is the shared handle that allows information to flow from one premises to the other and is then eliminated to produce a statement purely about S and P.

Mood, Figure, and Validity Classical syllogistic theory encodes statements as four preposition types (Table 8), with *mood* defined as the ordered triple across major premise, minor premise, and conclusion, and *figure* specifying the middle term's position, creating 256 possible forms of which only 24 are valid (Zong and Lin, 2024; Eisape et al., 2024; Copi et al., 2018). Validity requires: middle term distribution in at least one premise, no valid conclusion from two negative premises, and exactly one negative premise for negative conclusions (Hurley, 2014). Existential import enables subalternate moods like *Barbari* and *Darapti* (Parsons, 2014), making mood and figure as a compact notation and a complete decision procedure (Prior, 1962; Ramsey, 2009; Priest, 2008).

Trivial Validity Beyond the 24 structurally valid forms, some syllogistic arguments are formally valid for reasons that do not arise from the standard mood–figure interaction. These include *petitio principii*, where the conclusion merely restates a premise (Walton, 2008); immediate inferences such as valid conversion (restricted to E and I prepositions) and subalternation under existential import, where A entails I and E entails O (Hurley, 2014; Parsons, 2014); and cases in which contradictory premises trigger vacuous validity via the principle of explosion (*ex falso quodlibet*) (Priest, 2008).

3 System Overview

3.1 Normalization

3.1.1 Categorical Syllogism

The categorical syllogistic normalization process is defined as a transformation function $f : \mathcal{N} \rightarrow \mathcal{C}$, where $\mathcal{N}$ denotes the space of natural-language syllogistic arguments and $\mathcal{C}$ denotes the space of categorical syllogistic representations. Given an input argument $a \in \mathcal{N}$, the model first identifies exactly three distinct semantic categories $\{T_1, T_2, T_3\}$ corresponding to the subject, predicate, and middle term of the syllogism. These terms are then abstracted into symbolic constants $\{A, B, C\}$ according to their order of first appearance in the argument.

For example, consider the natural-language argument:

> **Premise 1:** Some housecats enjoy chasing mice.
> **Premise 2:** Any animal that enjoys chasing mice is a feline.
> **Conclusion:** All cats are animals.

The transformation function f extracts three terms and maps them as:

$$A : \text{animal}, \quad B : \text{feline}, \quad C : \text{cats}.$$

The argument is then normalized into standard categorical form:

$$\text{All } B \text{ are } A. \quad \text{All } C \text{ are } A. \quad \text{All } C \text{ are } B.$$

3.1.2 English Pivot Normalization

As most LLMs generally perform better in English (Guo et al., 2025), to handle languages beyond English, all non-English syllogisms are first processed through a constrained translation procedure using an LLM [1]. Instead of performing free-form translation, the model is instructed to extract the logical structure and translate only quantifiers and copular verbs into English. Furthermore, the original subject and predicate term is preserved in the the source language. This ensures structural standardization without introducing lexical drift that could alter term identity.

3.2 Preposition Parsing

After each data point is transformed into a canonical string of the form `P1. P2. Conclusion`, validity checking is extracted through a deterministic parsing procedure.

Each sentence P_i is first matched against a constrained set of regular-expression patterns and mapped to one of the four categorical types $f_i \in \{A, E, I, O\}$, while extracting its subject and predicate terms (s_i, p_i). Prior to matching, optional discourse markers (e.g., *therefore*, *thus*, *hence*) and minor surface variations (e.g., `is`/`are`) are normalized to ensure form consistency. This produces a structured representation:

$$\langle (f_1, s_1, p_1), (f_2, s_2, p_2), (f_3, s_3, p_3) \rangle.$$

---

[1] The detailed prompts used for assessing logical validity and premise relevance are shown in Appendix E

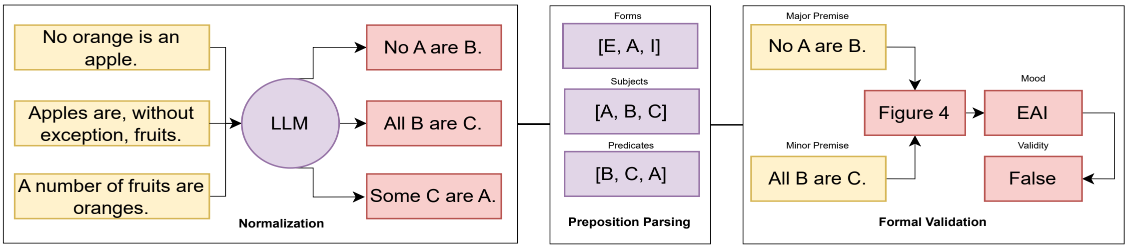

Figure 1: The flowchart illustrates the example step by step the flow of the proposed system.

Let $S = s_3$ and $P = p_3$ denote the subject and predicate of the conclusion. The middle term M is defined as the unique element in $(\{s_1, p_1\} \cap \{s_2, p_2\}) \setminus \{S, P\}$. The premise containing P is identified as the major premise, and the premise containing S as the minor premise. The figure is determined by the syntactic position (subject or predicate) of M within the major and minor premises, yielding one of the four canonical configurations. The mood is computed as the ordered triple $f_{\text{major}} f_{\text{minor}} f_{\text{conclusion}}$. The resulting $(\text{mood}, \text{figure})$ pair serves as the input to the subsequent formal validation step.

3.3 Formal Validation

3.3.1 Logical Validity

Given the parsed representation $\langle (f_1, s_1, p_1), (f_2, s_2, p_2), (f_3, s_3, p_3) \rangle$ and the inferred mood–figure pair (m, fig), logical validity is determined through a rule-based lookup procedure. For each figure $k \in \{1, 2, 3, 4\}$, we define a predefined set of valid moods $\mathcal{V}_k$. The syllogism is classified as valid if

$$\text{valid} = \nVdash\{m \in \mathcal{V}_{\text{fig}}\}.$$

We additionally detect trivially valid cases (e.g., a premise identical to the conclusion or valid E/I converses) to avoid misclassifying degenerate arguments as invalid. The implementation details are provided in Appendix B and C.

3.3.2 Relevant Premises Identification

For syllogisms classified as valid, the relevant-premise set $\mathcal{R} \subseteq \{1, 2\}$ consists of the two premises that structurally connect S and P through the middle term M. Concretely, the major premise is the premise containing P, and the minor premise is the premise containing S. Thus, $\mathcal{R}$ contains exactly those two premise indices. Meanwhile, if the syllogism is classified as invalid, we define $\mathcal{R} = \emptyset$ by convention.

Using the previously introduced example in Section 3.1.1, the conclusion connects $S = C$ and $P = A$. The premise containing P ("animal") serves as the major premise, while the premise containing S ("cats") serves as the minor premise. These two premises form the structurally sufficient pair that links S and P through the middle term M. Therefore, $\mathcal{R} = \{1, 2\}$, while any additional sentences, if present, are considered structurally irrelevant to the derivation.

4 Experimental Setup

Data Splits We use the official SemEval-2026 Task 11 data splits. The training set is used solely for prompt development and normalization strategy, while the development set is used for hyperparameter-free model comparison and ablation analysis. Final results are reported on the test set using the provided metric.

Normalization Model Premise normalization is performed using the Gemini 3 model, accessed via the official API.[2] The model is used to transform raw inputs (english-only and multilingual) into canonical syllogistic form.

Hyperparameter and Prompting All normalization prompts are fixed across splits. However, since the task includes english-only and multilingual premises, there are slight modifications to each task, following the experiment and ablation results. Furthermore, inference is performed with temperature = 0 and seed = 0 to ensure deterministic generation. No gradient-based fine-tuning is conducted.

Evaluation Metrics We report the official metrics defined in the shared task, including logical validity accuracy, Macro-averaged F1-Score for relevant premises, and combined score across english-only and multilingual tasks.

---

[2] We used the gemini-3-flash-preview

Normalization	Acc	Bias	Combined
Raw Data	92.10	10.63	26.66
PA Notation	77.48	8.17	24.08
FOL Notation	96.85	3.19	39.80
Syllogism Notation	**98.95**	**2.13**	**46.23**

Table 1: Performance comparison across different normalization strategies on english-only data.

5 Results

In this section, we discuss the findings across all four subtasks (See Table 2 for detailed results) . The discussion is divided into two parts: validity, which focuses on Subtasks 1 (English-only) and 3 (Multilingual), and relevant premise, which focuses on Subtasks 2 (English-only) and 4 (Multilingual).

5.1 Initial Experiment

We conducted an initial experiment to identify the most effective normalization strategy. We compared Predicate-Argument (PA Notation), First-Order Logic (FOL Notation), and Syllogism (Categorical Syllogism) and evaluated them based on the given metric. Overall, as shown in Table 1, normalization substantially reduces content-effect bias compared to raw inputs. Among the approaches, standard syllogistic notation achieved the highest combined score, indicating an optimal balance between abstraction and model interpretability. In contrast, fully formal symbolic representations (e.g., $\forall$, $\exists$, $\rightarrow$, $\neg$) resulted in lower combined performance, as increased parsing complexity outweighed the gains from bias reduction.

5.2 Validity Inference

5.2.1 English-only

We have managed to achieve perfect accuracy with zero bias in English logical validity. It shows that the normalization step accurately maps natural language premises to their formal quantifier-term representations. Therefore, the parsing-based approach guarantee correctness by construction due to well-defined validity conditions.

Conversely, the LLM-only setting decreased by approximately 2% in accuracy, resulting in four mismatches: two false positives and two false negatives. The two false positives indicate plausibility bias, where the model accepts conclusions that appear semantically reasonable despite not being logically entailed by the premises. The two false negatives, by contrast, reflect difficulty in handling partitive quantifiers such as “a number of,” suggesting sensitivity to linguistic variation rather than plausibility alone. These errors are non-deterministic, as repeated runs may yield different misclassifications, further showing that pattern-matching-based reasoning is less reliable than deterministic symbolic resolution.

Method	Acc	F1 (Premise)	Bias	Combined
Logical Validity (English)				
LLM-only	98.43	-	2.13	45.74
Norm + Parsing	**100**	-	**0.0**	**100**
Relevance Premises (English)				
LLM-only	95.78	**98.94**	5.0	34.87
Norm + Parsing	**98.94**	95.43	**2.0**	**46.31**
Logical Validity (Multilingual)				
LLM-only	96.87	-	4.16	36.66
Norm + Parsing	96.88	-	3.12	40.08
EPN + Norm + Parsing	**100**	-	**0.0**	**100**
Relevance Premises (Multilingual)				
LLM-only	86.98	87.76	7.29	28.05
Norm + Parsing	90.63	72.50	7.47	26.01
EPN + Norm + Parsing	**90.63**	**90.10**	**3.00**	**37.88**

Table 2: Comparison of different methods across subtasks. **LLM-only** denotes direct inference, **Norm + Parsing** is the deterministic parsing, and **EPN** refers to English Pivot Normalization.

5.2.2 Multilingual

In line with the perfect accuracy and zero bias achieved in the English-only setting, the approach extends effectively to the multilingual condition through an English-pivot normalization strategy. Empirically, omitting the translation step (Norm + Parsing) leads to a performance drop of over 3%, resulting in six structural mismatches. Error analysis indicates that, without translation, the cross-lingual variation in quantifier expression and term realization breaks the normalization step, producing unparseable structures that the deterministic rules cannot resolve.

Furthermore, the LLM-only baseline matches the accuracy of Norm + Parsing without translation, but with nearly double the bias score, and errors spanning six typologically diverse languages: Spanish, Swahili, Portuguese, Dutch, Bengali, and Russian. The false negatives mirror the English error patterns involving *E*-type premises, while the single false positive occurs in Bengali, suggesting additional difficulty with non-Latin scripts. The increased bias relative to the English setting confirms that LLM reasoning degrades on non-English input, reinforcing the advantage of the translate-first strategy.

5.3 Relevance Premises

5.3.1 English-only

The deterministic method (Norm + Parsing) performs strongly, achieving 98.94 accuracy and a 95.43 F1-score in English relevance premise identification. Although the LLM-only attains a slightly higher premise-level F1-score, 98.94, it is more susceptible to content effects, which ultimately reduces its overall combined score. Compared to validity inference, relevant-premise identification is inherently more challenging, as it requires selecting the structurally necessary premise. For example, minor representational overlaps or redundancies between premises can lead to prediction mismatches See Appendix D.2.1 for more details.

Out of 190 instances, 16 mismatches were observed in premise prediction. The remaining discrepancies largely stem from overlapping universal statements. For example, in one case, where both premises express universal relations involving the same middle term, the system selects the structurally sufficient premise connecting the predicate term and omits an additional universal statement. This is logically compatible but not strictly required under the structural criterion. These mismatches reflect representational redundancy rather than fundamental reasoning errors.

5.3.2 Multilingual

The improvements observed in the English-only setting largely persist in the multilingual evaluation. The EPN+Norm+Parsing and Norm+Parsing approaches achieve higher validity accuracy (90.63) than the LLM-only baseline (86.98). In contrast, the structured approaches derive validity deterministically from the identified mood and figure. Norm+Parsing, by comparison, constructs a locally coherent structure that maps to a valid mood even when the original argument is invalid, due to inadvertent incorporation of distractor premises. For premise selection, EPN+Norm+Parsing attains a higher F1 score (90.10) than Norm+Parsing (72.50), primarily because Norm+Parsing suffers from rigid singular–plural distinctions (e.g., *pianta/piante*, *rosa/rose*) that cause normalization failures during regular-expression matching.

The LLM-only model reasons holistically over the full multi-sentence input and is frequently distracted by irrelevant premises, leading it to miss the logically active pair, while EPN+Norm+Parsing may fail when a distractor sentence is selected as a premise. The LLM-only baseline frequently selects semantically related distractors, whereas EPN-based methods introduce errors when the LLM itself selects a distractor sentence as a premise during the EPN step, directly outputting an incorrect canonical form before any downstream processing occurs. See Appendix D.2.2 for more examples.

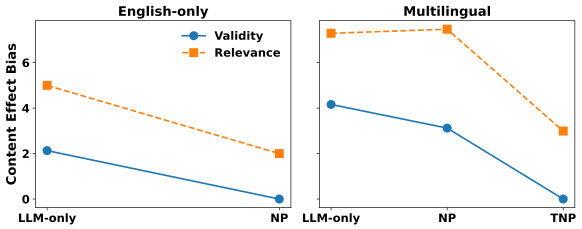

Figure 2: Content-effect reduction in English-only and Multilingual

5.4 Content-Effect Bias Reduction

Across both validity inference and premise identification, our method substantially reduces content-effect bias compared to LLM-only baselines (Figure 2). In the English-only setting, Norm+Parsing eliminates validity bias entirely and reduces relevance bias from 5.0 to 2.0. A similar reduction trend is observed in multilingual evaluation: validity bias decreases from 4.16 (LLM-only) to 3.12 (Norm+Parsing) and to 0.0 under ENP+Norm+Parsing, while relevance bias drops markedly from 7.29 (LLM-only) and 7.47 (Norm+Parsing) to 2.99 with translation-based normalization. These consistent reductions highlight our hypothesis that formal syllogistic structure effectively mitigates content-effect biases across tasks and languages. By abstracting world-specific lexical through normalization and retaining only formal syllogistic structure, the deterministic method directly targets this source of interference.

6 Conclusion

By transforming syllogisms into canonical representations and applying deterministic parsing, our structural abstraction method achieves top-5 rankings across all SemEval-2026 Task 11 subtasks while substantially reducing content effects. This simple approach offers a compelling alternative to complex architectural modifications and opens promising avenues for developing scalable, interpretable reasoning techniques across languages.

Acknowledgments

We sincerely thank Onno Kampman from SEACrowd for generously providing Google Gemini credits, which made the execution of our experiments possible.

Limitations

This paper evaluates only one commercial model, Gemini-3-Flash, and does not explore other commercial or open-source models. In addition, we use a fully deterministic decoding strategy (greedy decoding with temperature set to 0 and a fixed seed). We do not conduct experiments across multiple random seeds or sampling settings, and therefore do not examine performance variability or leverage the potential diversity of large language models.

Ethical Considerations

We acknowledge that our research utilized AI tools for writing, rewriting, and generating code. Although these tools offer significant advantages in terms of efficiency and productivity, their use raises important ethical considerations. We recognize the potential for bias and errors inherent in AI-generated content and have taken steps to mitigate these risks through rigorous human review and validation. Furthermore, we are mindful of the potential impact on the broader software development community, particularly regarding job displacement and the need for upskilling. We believe that responsible AI integration should prioritize transparency, accountability, and the empowerment of human developers, ensuring that these tools augment rather than replace human expertise. This research aims to contribute to the ongoing dialogue on ethical AI development and usage, advocating for a future where AI tools are harnessed responsibly to enhance human creativity and innovation in the field of software engineering.

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, Miami, Florida, USA. Association for Computational Linguistics.

Irving M. Copi, Carl Cohen, and Victor Rodych. 2018. *Introduction to Logic*, 15th edition. Routledge, New York.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893. Association for Computational Linguistics.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *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.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. Association for Computational Linguistics.

Patrick J. Hurley. 2014. *A Concise Introduction to Logic*, 12th edition. Cengage Learning, Stamford, CT.

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, Vienna, Austria. Association for Computational Linguistics.

Joanito Agili Lopo, Muhammad Ravi Shulthan Habibi, Tack Hwa Wong, Muhammad Ilham Ghozali, Fajri Koto, Genta Indra Winata, Peerat Limkonchotiwat, Alham Fikri Aji, and Samuel Cahyawijaya. 2025. Language surgery in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 438–467, Suzhuo, China. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

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.

Terence Parsons. 2014. *Articulating Medieval Logic*. Oxford University Press, Oxford.

Graham Priest. 2008. *An Introduction to Non-Classical Logic: From If to Is*, 2nd edition. Cambridge University Press, Cambridge.

Arthur N. Prior. 1962. *Formal Logic*, 2nd edition. Oxford University Press.

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, Miami, Florida, USA. Association for Computational Linguistics.

Frank P. Ramsey. 2009. On a problem of formal logic. In Ira Gessel and Gian-Carlo Rota, editors, *Classic Papers in Combinatorics*, Modern Birkhäuser Classics. Birkhäuser Boston.

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, Vienna, Austria. 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. *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)*.

Douglas Walton. 2008. *Informal Logic: A Pragmatic Approach*, 2nd edition. Cambridge University Press, Cambridge.

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.

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, Bangkok, Thailand. Association for Computational Linguistics.

Shi Zong and Jimmy Lin. 2024. Categorical syllogisms revisited: A review of the logical reasoning abilities of LLMs for analyzing categorical syllogisms. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 230–239. Association for Computational Linguistics.

A Leaderboard Comparison

Below we present the top-5 comparison results for each subtask: Table 3 corresponds to Subtask 1, Table 4 to Subtask 2, Table 5 to Subtask 3, and Table 6 to Subtask 4. Our team is registered as **itlc_team** on Codabench; for simplicity, we refer to it as **ITLC**. For Subtask 2 (Table 4), our submission appears under the name **joanitolopo**, due to changes in the test data made by the organizers during the evaluation phase. In addition, the scores on the leaderboard differ from those reported in the current paper, particularly for Subtask 4, since a unified prompt was used across all subtasks in this paper. However, the overall ranking remains the same.

#	Team Name	Accuracy	Content Effect	Combined
1	rongchuan	100	0.0	100
...	...	...	...	...
9	ITLC	100	0.0	100
...	...	...	...	...
12	ewelinaksiez	97.91	0.02	95.81
13	chisnguyen	99.48	1.06	57.68
14	vinaybabu	99.48	1.06	57.68

Table 3: Leaderboard comparison of Subtask 1

#	Team Name	Accuracy	F1	Bias	Combined
1	Habib_TAZ	100	100	0.0	100
2	YNU-NLP	100	100	0.0	100
3	PA	100	98.95	0.0	99.47
4	junhaofu	96.84	94.21	1.13	54.43
5	butasrafael	96.88	95.83	1.17	54.24
6	ITLC	98.94	95.43	2.0	46.31

Table 4: Leaderboard comparison of Subtask 2

#	Team Name	Accuracy	Content Effect	Combined
1	Habib_TAZ	100	0.0	100
2	PA	100	0.0	100
3	ITLC	100	0.0	100
4	junhaofu	95.83	0.17	82.59
5	rongchuan	96.35	1.0	56.97

Table 5: Leaderboard comparison of Subtask 3

#	Team Name	Accuracy	F1	Bias	Combined
1	PA	100	94.36	0.0	97.18
2	YNU-NLP	90.1	89.58	1.26	49.51
3	sungbin_kai	86.46	86.8	1.2	48.48
4	ufal_cuni	84.9	83.42	1.37	45.2
5	ITLC	91.15	98.31	2.19	43.83

Table 6: Leaderboard comparison of Subtask 4

B Parsing algorithm

Algorithm 1 Parse one preposition into (form, subj, pred)

```
function MatchAEIO($p$)
    $p \leftarrow$ lowercase($p$); trim($p$);
    replace “ is ” with “ are ”
    remove leading connector in {therefore, thus, hence, so}
    if $p$ matches “all $X$ are $Y$” then
        return $(A, X, Y)$
    end if
    if $p$ matches “no $X$ are $Y$” then
        return $(E, X, Y)$
    end if
    if $p$ matches “some $X$ are not $Y$” then
        return $(O, X, Y)$
    end if
    if $p$ matches “some $X$ are $Y$” then
        return $(I, X, Y)$
    end if
    return fail
end function
```

Algorithm 2 Infer mood/figure and validate by lookup

```
function ValidateSyllogism($x$)
    $parts \leftarrow$ SplitNonEmpty($x$,".")
    if $|parts| \neq 3$ then
        return $(\emptyset, 0, \textbf{false})$
    end if
    $(p_1, p_2, c) \leftarrow (parts[1], parts[2], parts[3])$
    $(f_1, s_1, r_1, ok_1) \leftarrow$MatchAEIO($p_1$)
    $(f_2, s_2, r_2, ok_2) \leftarrow$MatchAEIO($p_2$)
    $(f_3, s_3, r_3, ok_3) \leftarrow$MatchAEIO($c$)
    if not $(ok_1 \wedge ok_2 \wedge ok_3)$ then
        return $(\emptyset, 0, \textbf{false})$
    end if
    $S \leftarrow s_3,\ P \leftarrow r_3,\ U_1 \leftarrow \{s_1, r_1\},\ U_2 \leftarrow \{s_2, r_2\}$
    if $|U_1 \cup U_2 \cup \{S, P\}| \neq 3$ then
        return $(\emptyset, 0, \textbf{false})$
    end if
    $Mset \leftarrow (U_1 \cap U_2) \setminus \{S, P\}$
    if $|Mset| \neq 1$ then
        return $(\emptyset, 0, \textbf{false})$
    end if
    $M \leftarrow$Only($Mset$)
    if $P \in U_1$ then
        $maj \leftarrow 1$; $min \leftarrow 2$
    else if $P \in U_2$ then
        $maj \leftarrow 2$; $min \leftarrow 1$
    else
        return $(\emptyset, 0, \textbf{false})$
    end if
    $a \leftarrow (s_{maj} = M)$; $b \leftarrow (s_{min} = M)$
    if $a \wedge \neg b$ then
        $figure \leftarrow 1$
    else if $\neg a \wedge \neg b$ then
        $figure \leftarrow 2$
    else if $a \wedge b$ then
        $figure \leftarrow 3$
    else
        $figure \leftarrow 4$
    end if
    $mood \leftarrow (f_{maj}, f_{min}, f_3)$
    return $(mood, figure,\ mood \in VALID[figure])$
end function
```

C Lookup Table

Table 7: Lookup table for valid moods by figure

Figure	Valid moods (major–minor–conclusion)
1	AAA, EAE, AII, EIO, AAI, EAO
2	EAE, AEE, EIO, AOO, EAO, AEO
3	AAI, IAI, AII, EAO, OAO, EIO
4	AAI, AEE, IAI, EAO, EIO, AEO

D Example Appendix

D.1 Logical Validity Premises

D.1.1 English only

As discussed in Section 5.2.1, our normalization approach achieved 100% accuracy due to the deterministic nature of preposition parsing. In contrast, Table 9 illustrates a representative LLM-only error where the model predicts the argument as valid by relying on semantic plausibility rather than strict logical entailment. The conclusion “some vehicles are bikes” is true under general world knowledge, and this real-world truthfulness appears to bias the model toward accepting it. However, the conclusion is not derivable from the given premises: an E-type premise (“No bikes are cars”) paired with an A-type premise (“All bikes are vehicles”) does not license an existential affirmative conclusion about vehicles being bikes without assuming existential import. The deterministic parser correctly rejects this form because no valid mood-figure combination produces such a conclusion from the given premise types, regardless of whether the conclusion happens to be true in the real world. This distinction between semantic truth and logical validity

Table 8: Four types of categorical sentences and their translations into predicate-logic and set-theoretic notation.

Type	Sentence Pattern	Predicate Logic	Set Theory	Description
A (all)	All S are P	$\forall x\,(S(x) \rightarrow P(x))$	$S \subseteq P$	Universal Affirmative
E (no)	No S are P	$\forall x\,(S(x) \rightarrow \neg P(x))$	$S \cap P = \emptyset$	Universal Negative
I (some)	Some S are P	$\exists x\,(S(x) \wedge P(x))$	$S \cap P \neq \emptyset$	Particular Affirmative
O (some-not)	Some S are not P	$\exists x\,(S(x) \wedge \neg P(x))$	$S \setminus P \neq \emptyset$	Particular Negative

is precisely where LLM-based inference breaks down, as the model conflates what is plausible with what is entailed.

D.1.2 Multilingual

The EPN + Norm + Parsing pipeline achieves perfect accuracy on multilingual validity as shown in Section 5.2.2, confirming that the deterministic approach generalizes fully across languages when EPN is applied. However, when relying on the LLM-only baseline, this guarantee breaks down. Table 10 illustrates a representative failure. The Spanish construction “no se trata de que cada” is a negated universal, which under standard categorical logic maps to a particular negative (O-type: “Some S are not P”). The LLM instead misinterprets the surface-level negation marker “no” as indicating a universal negative (E-type: “No S are P”), collapsing the distinction between sentential negation and quantifier negation. This misclassification propagates through the inference chain, causing the model to reject a valid conclusion. The error is particularly revealing because the underlying syllogistic structure is straightforward once the premises are correctly typed. The EPN-first pipeline avoids this entirely: translation preserves the negated-universal semantics in English (“it is not the case that every...”), and the normalization step deterministically maps this pattern to O-type before symbolic resolution is applied. This confirms that the bottleneck for multilingual validity is not logical reasoning itself, but quantifier interpretation across languages.

Table 11 illustrates the primary failure mode of the Norm + Parsing pipeline when applied without prior translation. The original English syllogism contains three distinct terms (“dog,” “poodle,” “canine”) and is straightforwardly valid. However, French lacks a lexical distinction between “dog” and “canine,” collapsing both into “chien.” This translation-induced term collapse reduces the second premise from a meaningful categorical statement (“All dogs are canines”) to a tautology (“All chiens are chiens”), producing a degenerate two-term structure that the parser assigns Figure 0 with undefined mood. The deterministic rules therefore reject the argument despite the original being logically sound. This pattern recurs across five of the six errors in the Norm + Parsing setting, where cross-lingual lexical gaps or synonym merging similarly destroy the term structure required for symbolic resolution. The EPN + Norm + Parsing pipeline avoids this entirely by operating on the English source, where the three-term distinction is preserved by construction.

D.2 Relevance Premises

D.2.1 English-only

Table 12 illustrates the example of minor representational overlaps or redundancies between premises. Table 13 shows that LLM failed to select any relevant premises, while Table 14 shows false positive selection example by LLM.

D.2.2 Multilingual

Table 15 illustrates a case where the LLM-only approach fails to identify the active premise pair due to semantically plausible distractors; Table 16 shows how Norm+Parsing constructs a locally coherent valid-looking structure from an invalid argument; and Table 17 demonstrates a case where the EPN step itself selects a distractor as a premise, directly producing an incorrect canonical form before any downstream processing.

E Prompt Appendix

Figure 3 shows the LLM-only prompt used across all subtasks, Figure 4 shows the normalization prompt used across all subtasks, Figure 5 shows the EPN prompt used in Subtask 3, Figure 6 shows the EPN prompt used in Subtask 4, and Figure 7 shows its variant that incorporates the Google-translated sentence.

F Google Translation as a Double-Edged Sword in Multilingual Settings

This section examines the impact of Google-translated sentences in the Relevance Premise

Idx	Content	Interpretation	Role	Inference Issue
0	There are no bikes that can be called cars.	No bikes are cars.	premise	–
1	It is also true that every bike is a type of vehicle.	All bikes are vehicles.	premise	–
2	This has led to the conclusion that a portion of vehicles are bikes.	Some vehicles are bikes.	conclusion	Semantically plausible, but not logically entailed.
GT validity: **False** LLM-only prediction: **True**				

Table 9: LLM-only (Subtask 1) exhibits plausibility bias by accepting a semantically plausible conclusion that is not logically entailed by the premises.

Idx	Content	Post EPN	Interpretation	Role	Error Note
0	No se trata de que cada llave sea un objeto.	It is not the case that every key is an object.	Some keys are not objects. (O)	premise	–
1	Todo lo que sea una llave inglesa es una herramienta.	Everything that is a wrench is a tool.	All wrenches are tools. (A)	premise	–
2	Por tanto, se puede concluir que algunas herramientas no son objetos.	Therefore, some tools are not objects.	Some tools are not objects. (O)	conclusion	LLM fails to convert the negated universal ("no se trata de que cada") into O-type; treats it as E-type, breaking the inference chain.
GT validity: **True** LLM-only prediction: **False** Language: **Spanish**					

Table 10: LLM-only (Subtask 3) fails to resolve a negated universal quantifier in Spanish, misinterpreting "no se trata de que cada" as a universal negative rather than the intended particular negative.

(Multilingual) setting, which refer to subtask 4. We replace original sentences with their Google translations in Norm+Parsing, and provide translated sentences as references to the LLM in EPN+Norm+Parsing. Due to linguistic normalization, EPN often collapses synonyms and singular–plural distinctions into a single English form. This generally improves premise F1 but can harm validity accuracy, making translation a double-edged sword.

F.1 Impact on Relevance Premise Selection

As shown in Table 20, Google Translate collapses *piante/pianta* and *rosa/rose* into the same English terms (*plant*, *rose*). In Norm+Parsing, this enables successful matching using regular expressions and increases premise F1 from 72.50 to 88.39 (Table 18).

However, in EPN+Norm+Parsing, premise selection relies on LLM reasoning rather than regular expressions. Consequently, EPN slightly reduces F1 (90.10 to 89.58), as term collapse can distort the reasoning signal.

F.2 Impact on Validity

As illustrated in Table 19, EPN collapses *cachorro* and *cães* into *dog*, making the chain (*canines* → *dog* → *mammal*) more transparent in English. This biases the LLM toward selecting an incorrect premise pair, producing a spurious validity form.

Without EPN, the LLM explicitly treats *cachorro* and *cães* as distinct surface forms and avoids merging them. As a result, translation reduces validity accuracy and increases content bias: in Norm+Parsing, accuracy drops from 90.63 to 88.54 and bias rises from 7.47 to 8.14; in Translate+Norm+Parsing, accuracy decreases from 90.63 to 89.58 and bias increases from 2.99 to 4.32, as shown in Table 18.

Idx	Original (English)	Content (French)	Interpretation	Role	Error Note
0	It is not true that every dog is a poodle.	Ce n'est pas vrai que tous les chiens sont des caniches.	Some dogs are not poodles. (O)	premise	–
1	Every creature that is a dog is a **canine**.	Toute créature qui est un **chien** est un **chien**.	All A are A. (tautology)	premise	French collapses "canine" into "chien" (dog), destroying the three-term structure.
2	Some **canines** are not poodles.	Certains **chiens** ne sont pas des caniches.	Some dogs are not poodles. (O)	conclusion	–

GT validity: **True**   Norm + Parsing prediction: **False**   Language: **French**

Table 11: Norm + Parsing without translation (Subtask 3) fails on a French syllogism where the translation collapses "canine" and "dog" into the same French word "chien," reducing a valid three-term syllogism to a degenerate two-term structure.

Idx	Premise	Role	Prediction
0	There are no circles that are also three-sided figures.	Relevant	Selected
1	Some isosceles triangles are three-sided figures.	Distractor	Wrongly selected
2	All scalene triangles are three-sided figures.	Distractor	–
3	Every equilateral triangle is a three-sided figure.	Distractor	–
4	All figures that are triangles are three-sided figures.	Relevant	Missed
5	Circles are not triangles.	Distractor	–

Gold Relevant Premises: [0, 4]   **Predicted Relevant Premises:** [0, 1]

Observation: Premise (1) shares the surface concept *three-sided figures* with the gold premises, creating a minor representational overlap that attracts selection

Table 12: Example of minor representational overlap in Subtask 4. The model selects premise (1) due to shared surface terminology ("three-sided figures"), although premise (4) is required for the correct logical structure.

Idx	Premise	Role	Prediction
0	Some butterflies have colorful wings.	Distractor	–
1	It is true that all beetles have six legs.	Distractor	–
2	There are no webs spun by grasshoppers.	Relevant	Missed
3	Spiders are never ants.	Relevant	Missed
4	Every single bee pollinates flowers.	Distractor	–
5	A few ants are, in fact, insects.	Distractor	–
6	There exist insects that are not spiders.	Distractor	–

Gold Relevant Premises: [2, 3]   **Predicted Relevant Premises:** []

Observation: Instead of tracking the negation relations required for the valid inference, LLM overlooks the necessary premises entirely, leading to an incorrect validity prediction.

Table 13: Example of LLM-only relevant premises mismatch. Although premises (2) and (3) are required to establish the contradiction structure, the model fails to select any relevant premises and predicts an incorrect validity label.

Idx	Premise	Role	Prediction
0	Some hounds are actually birds.	Distractor	–
1	Any retriever is a type of fish.	Distractor	–
2	Anything that is a poodle is also a canine.	Distractor	Wrongly selected
3	There are no animals that bark.	Distractor	–
4	It is a fact that all collies have scales.	Distractor	–
5	Every single puppy is a kitten.	Distractor	–
6	The entire set of poodles is contained within the set of dogs.	Distractor	–
7	A portion of dogs are not canines.	Distractor	Wrongly selected

Gold Relevant Premises: [] **Predicted Relevant Premises:** [2, 7]

Observation: The model selects premises (2) and (7) due to lexical and conceptual overlap between *poodles, dogs,* and *canines*, forming an apparent contradiction. However, these premises are structurally irrelevant to the target conclusion, resulting in a false-positive relevance attribution and incorrect validity judgment.

Table 14: Example of false-positive premise selection. Although no premise is structurally required for the invalid syllogism, the model selects (2) and (7) due to surface-level semantic overlap, leading to an incorrect validity prediction.

Idx	Original (IT)	English	Role
0	Tutto ciò che è un uccello depone le uova	Anything that is a bird lays eggs	GT premise
1	Ogni singola anatra è un uccello	Every single duck is a bird	Distractor
2	Alcuni uccelli sono creature che costruiscono nidi	Some birds are creatures that build nests	Distractor
3	Qualsiasi struzzo è un uccello che depone le uova	Any ostrich is a bird that lays eggs	Distractor
4	Non esistono pinguini che non depongono le uova	There are no penguins that do not lay eggs	Distractor
5	Non esistono polli che non siano uccelli	There are no chickens that are not birds	GT premise
6	È il caso che alcune galline depongano le uova	It is the case that some chickens lay eggs	Conclusion

GT premises: [0, 5] LLM-only prediction: **[]** (predicted invalid)

Table 15: LLM-only (subtask 4) misses the active premise pair [0,5]; semantically plausible distractors about ostriches and penguins lead the model to predict invalid.

Idx	Original (ES)	English	Role
0	Cada brócoli es una verdura	Every single broccoli is a vegetable	Distractor
1	Es cierto que todas las frutas son comestibles	It is true that all fruits are edible	Distractor
2	Cualquier patata es una verdura	Any potato is a vegetable	Distractor
3	Algunas cosas comestibles son granos	Some edible things are grains	Distractor
4	Hay algunas raíces que son comestibles	There are a few roots that are edible	Distractor
5	Cada vegetal es comestible	Every single vegetable is edible	Wrongly selected as P1
6	Todo lo que es una zanahoria es una verdura	Anything that is a carrot is a vegetable	Wrongly selected as P2
7	No hay zanahorias que no sean comestibles	There are no carrots that are not edible	Conclusion

Norm output: *"All verdura are comestible. All zanahoria are verdura.Therefore, All zanahoria are comestible."*

GT Validity: False Norm+Parsing Prediction: **True**

Table 16: Norm+Parsing (subtask 4) selects sentences [5,6] forming a plausible carrot→vegetable→edible chain that maps to a valid mood, despite the overall argument being invalid.

Idx	Original (FR)	English	Role
0	Chaque chiot naît en sachant parler français	Every single puppy is born knowing how to speak French	Distractor
1	Aucun chien n'a jamais marché sur quatre pattes	No canine has ever walked on four legs	Distractor
2	Il n'existe pas d'animaux qui soient aussi des chiens	There exist no animals that are also canines	GT premise
3	Il est indéniable que le ciel est vert et l'herbe bleue	The sky is green and the grass is blue	Distractor
4	Tous les chiens sont des robots déguisés	All dogs are actually disguised robots from space	Distractor
5	Certains animaux sont entièrement en verre	Some animals are made entirely of glass	Wrongly selected as P1
6	L'ensemble des chiens fait partie des canines	The entire set of dogs is a part of the canines	GT premise
7	Il n'y a pas d'animaux qui soient des chiens	There are no animals which are dogs	Conclusion

EPN output: *"Some animaux are entièrement en verre. All chiens are canines. Therefore, No animaux are chiens."*

GT premises: [2, 6] EPN+Norm+Parsing Prediction: **[]** (predicted invalid)

Table 17: The EPN step (subtask 4) selects sentence [5] (a distractor) as P1 instead of the correct premise [2], producing a structure that the downstream classifier rejects as invalid.

Method	Acc	F1 (premise)	Bias	Combined
Norm + Parsing	90.63	72.50	7.47	26.01
Norm + Parsing using google translate sentence	88.54	88.39	8.14	27.54
EPN + Norm + Parsing	**90.63**	**90.10**	3.00	**37.88**
+ google translate example	89.58	89.58	4.32	33.53

Table 18: Ablation study on the impact of incorporating Google-translated sentences in a multilingual setting

	S1	S5	S7	Conclusion
Original (PT)	Alguns mamíferos não são cães	É certo que todo *cachorro* é um mamífero	Todos os caninos são *cães*	Todos os caninos são mamíferos
Google Trans.	Some mammals are not dogs	It is true that every **dog** is a **mammal**	All **canines** are **dogs**	All **canines** are **mammals**
Ground Truth	Some **mammals** are not **dogs**	Every single dog is a mammal	All **canines** are **dogs**	All **canines** are **mammals**

Table 19: Effect of lexical collapse in the Google-translated reference. Highlighted cells show selected premises per system. Terms are coloured consistently: **canines**, **mammals**, **dogs**.

LLM-only prompt

```
You are a strict classical categorical logic evaluator.

You are given:
- An ID
- A syllogism containing multiple premises and one conclusion.

----------------------------------------
INSTRUCTIONS

1. All sentences before the final "Therefore" are premises.
   The sentence after "Therefore" is the conclusion.

2. Internally rewrite each statement into standard categorical form:
   - All X are Y (A)
   - No X are Y (E)
   - Some X are Y (I)
   - Some X are not Y (O)

   Handle paraphrases such as:
   - Every X is Y
   - Not a single X is Y
   - At least one X is not Y
   - Double negations

3. Determine whether the conclusion NECESSARILY follows
   from a subset of the premises under classical categorical logic.

4. If the argument is valid:
   - Identify the MINIMAL set of premises required to entail the conclusion.
   - Return their indexes (0-based).
   - Indexing is based on order of appearance in the text.
   - Do NOT include unused premises.

5. If the argument is invalid:
   - validity = false
   - relevant_premises = []

6. Do NOT explain reasoning.
   Output JSON ONLY.

STRICT REQUIREMENTS:
- Output must be valid JSON.
- No explanation.
- No markdown.
- No extra keys.
- Only "validity" and "relevant_premises".

----------------------------------------
OUTPUT FORMAT:

{{
  "validity": true or false,
  "relevant_premises": [int, int]
}}

----------------------------------------
SYLLOGISM:
{syllogism}
```

Figure 3: LLM-only prompt for retrieve the validity and relevant premise directly

	P1	P2	Conclusion
Original (IT)	Tutto ciò che è una **rosa** è anche una **pianta**	Esistono **rose** che sono **fiori**	Alcuni **fiori** sono **piante**
Google Trans.	Everything that is a **rose** is also a **plant**	There are **roses** that are **flowers**	Some **flowers** are **plants**

Table 20: Italian syllogism with Google-translated reference. Terms are coloured consistently: **rose**, **flowers**, **plants**.

Normalization prompt

```
Transform a syllogistic argument into symbolic notation.

STEP 1 - IDENTIFY THE 3 TERMS:
- Find exactly 3 distinct categories/terms
- Middle term (M) = appears in BOTH premises, NEVER in conclusion
- Subject of conclusion (S) = subject in conclusion + appears in one premise
- Predicate of conclusion (P) = predicate in conclusion + appears in one premise

STEP 2 - MAP TERMS TO LETTERS:
- Assign A, B, C by order of first appearance in text

STEP 3 - CONVERT TO STANDARD FORM:
- "All X are Y" / "Every X is Y" / "X is subset of Y" / "Anything that is X is Y" → All X are Y (A)
- "No X are Y" / "Not a single X is Y" / "X cannot be Y" / "X is never Y" → No X are Y (E)
- "Some X are Y" / "A few X are Y" / "There exist X that are Y" / "Something that is X is Y" → Some X are Y (I)
- "Some X are not Y" / "Not all X are Y" / "At least one X is not Y" → Some X are not Y (O)
- "No X are not Y" / "There are no X that are not Y" → All X are Y (A) [double negative]

OUTPUT FORMAT (JSON only, no markdown):
{{
  "reasoning": "<free-form step-by-step analysis>",
  "mapped": "A:<term1>,B:<term2>,C:<term3>",
  "parsed": "<P1>. <P2>. <conclusion>"
}}

Transform:
{syllogism}
```

Figure 4: Norm prompt for normalize sentences into standard categorical form

EPN prompt for Validity Inference , Multilingual Setting (Subtask 3)

```
You are an expert translator for categorical syllogisms.

TASK
Given {syllogism}, extract its logical structure.
Translate quantifiers and verbs into English. Keep ONLY subject and predicate terms in the SOURCE language.

RULES:
- Output format per sentence: [Quantifier] [native_subj] are [native_pred]
- Quantifiers: All / No / Some / Some...not
- Never produce "All X are not Y".
- Preserve source subject as subject, source predicate as predicate. Do NOT swap.
- Remove rhetorical wrappers.

QUANTIFIER GUIDE:
- "every/all/any/anything that is X/each/whatever is" → All
- "no/none/never/there are no/not a single" → No
- "some/a few/certain/there are some/a portion" → Some
- "not all/not every" → Some...not
- "it is not the case that X are Y" with bare plurals → No.
- "anything that is X is Y" → All.

TERM DISTRIBUTION:
After extracting subj and pred for all 3 sentences, count how many sentences
each distinct native term appears in. Each term MUST appear in exactly 2 of 3 sentences.
Treat singular/plural variants as the same term (e.g. gari/magari).

If a term appears in 3 sentences and another appears in only 1:
CHECK: is the 1-count term a BROADER or NARROWER category of the 3-count term?
  - YES (e.g. maglietta/camicie = t-shirts/shirts): REPLACE one occurrence so both = 2.
  - NO (e.g. scrittori/individuo che scrive libri = writers/person who writes books,
    where one is a DEFINITION or DESCRIPTION, not a category): DO NOT replace.
    The 3-count term legitimately appears in all 3 sentences. Leave as-is.

If a term fills both slots in one sentence AND no other term covers its second sense:
tag it word[s] (specific) and word[g] (general).
Analyze verb context: relative clause ("thing that is X") = [s], main predicate = [g].
Propagate so each tagged term is in exactly 2 sentences.
Use the SAME base form for all tagged occurrences (pick one, use it everywhere).
Note: "poly: word → specific=[english], general=[english]".



REASONING FORMAT:
1. Extract subj/pred per sentence
2. Count term distribution
3. Fix distribution if needed (replace or tag)
4. Output

OUTPUT (JSON only, no markdown):
{{
  "detected_language": "<lang>",
  "reasoning": "<extract, count, fix, output>",
  "english": "P1. P2. Therefore, C."
}}

EXAMPLES:
Input: "All goyangi are dongmul. Some dongmul are not poyuryu. Therefore, some goyangi are not poyuryu."
reasoning: "Extract: P1 subj=goyangi pred=dongmul. P2 subj=dongmul pred=poyuryu. C subj=goyangi pred=poyuryu.
Count: goyangi=2, dongmul=2, poyuryu=2."
english: "All goyangi are dongmul. Some dongmul are not poyuryu. Therefore, some goyangi are not poyuryu."

Replacement needed (3-count and 1-count):
Input: "Todos los autos son máquinas. Hay carros que son carros. Algunos carros son máquinas."
reasoning: "Extract: P1 subj=autos pred=máquinas. P2 subj=carros pred=carros. C subj=carros pred=máquinas.
Count: autos=1, máquinas=2, carros=3. Fix: replace P2 pred carros with autos.
Count: autos=2, máquinas=2, carros=2. "
english: "All autos are máquinas. Some carros are autos. Therefore, Some carros are máquinas."

Replacement across sentences:
Input: "Nenhum gato é um pássaro. Todo gato é um animal doméstico. Pássaros não são gatos."
reasoning: "Extract: P1 subj=gato pred=pássaro. P2 subj=gato pred=animal doméstico. C subj=pássaro pred=gato.
Count: gato=3, pássaro=2, animal doméstico=1. Fix: replace P1 subj gato with animal doméstico. Wait — that changes
the source subject.
Instead: C says 'pássaros não são gatos' but gato=3, animal doméstico=1. They share a meaning. Replace C pred gato
with animal doméstico.
Count: gato=2, pássaro=2, animal doméstico=2. "
english: "No gato are pássaro. All gato are animal doméstico. Therefore, No pássaro are animal doméstico."

Tagging needed (same word both slots, no other term):
Input: "Hakuna samaki ni nyoka. Kila nyoka ni nyoka. Nyoka fulani si samaki."
reasoning: "Extract: P1 subj=samaki pred=nyoka. P2 subj=nyoka pred=nyoka. C subj=nyoka pred=samaki.
Count: samaki=2, nyoka=4. No 1-count term to replace. P2 has same word both slots. Tag: P2 subj=nyoka[s] pred=nyoka[g].
Propagate: nyoka[s] in P2,C. nyoka[g] in P1,P2. samaki in P1,C. Each=2. . poly: nyoka → specific=snakes, general=reptiles."
english: "No samaki are nyoka[g]. All nyoka[s] are nyoka[g]. Some nyoka[s] are not samaki."

Polysemy with verb context:
Input: "Nicht alle Tiere sind Hunde. Jedes Geschöpf, das ein Tier ist, ist ein Tier. Daher sind einige Tiere keine Hunde."
reasoning: "Extract: P1 subj=Tiere pred=Hunde. P2 subj=Tier pred=Tier. C subj=Tiere pred=Hunde. Count: Tiere/Tier=4, Hunde=2.
No 1-count term. P2 same word both slots. Tag using verb context: 'Geschöpf, das ein Tier ist' relative clause = [s], 'ist ein Tier'
predicate = [g]. P2 subj=Tier[s] pred=Tier[g]. Propagate: Tier[s] in P1,P2. Tier[g] in P2,C. Hunde in P1,C. Each=2. .
poly: Tier → specific=animals, general=creatures."
english: "Some Tiere[s] are not Hunde. All Tier[s] are Tier[g]. Therefore, Some Tiere[g] are not Hunde."
```

Figure 5: EPN prompt for Subtask 3 for extract subject term

EPN prompt for Relevance Premises , Multilingual Setting (Subtask 4)

```
You are an expert logical translator for categorical syllogisms.

TASK
Given {syllogism}, extract exactly:
- Two premises
- One conclusion

Preserve the original logical structure.
Translate ONLY quantifiers and copula into English.
Keep subject and predicate terms EXACTLY as in the source language.

STRICT RULES:

1. FORMAT
Each sentence must follow:
[Quantifier] [native_subject] are [native_predicate]

Allowed quantifiers:
All / No / Some / Some...not

Never output:
"All X are not Y"

2. DO NOT MODIFY LOGIC
- Do NOT swap subject and predicate.
- Do NOT replace terms.
- Do NOT merge synonyms.
- Do NOT normalize or repair the argument.
- Do NOT balance term distribution.
- Do NOT split polysemous words.
- Do NOT introduce [s] or [g] tags.

Extract the argument exactly as written.

3. TERM HANDLING
- Copy subject and predicate verbatim.
- Singular/plural variants count as the same term.
- Descriptive phrases remain descriptive phrases.
- Identity statements (X are X) remain unchanged.

4. TERM COUNT CHECK (Diagnostic Only)
After extraction:
- Count distinct terms.
- A classical syllogism should have exactly 3 distinct terms.
- If not, DO NOT repair.
- Simply report the count in reasoning.

5. SENTENCE SELECTION
If more than 3 sentences are present:
- Select the two premises and the conclusion that form the main argument.
- Ignore rhetorical commentary.

REASONING FORMAT:
1. Identify selected sentences (P1, P2, C)
2. Extract subject and predicate
3. Count distinct terms (no fixing)
4. Output final structured form

OUTPUT (JSON only, no markdown):
{
  "detected_language": "<lang>",
  "reasoning": "<selection + extraction + term_count>",
  "english": "P1. P2. Therefore, C."
}
```

Figure 6: EPN prompt for Subtask 4 for extract and filter relevant premise and conclusion

EPN prompt with google translated sentence for Relevance Premises , Multilingual Setting (Subtask 4)

```
You are an expert logical translator for categorical syllogisms.

TASK
Given Original syllogism in source language:{syllogism},
Google Translation reference:{google_translated_Sentence}
extract exactly:
- Two premises
- One conclusion

Preserve the original logical structure.
Translate ONLY quantifiers and copula into English.
Keep subject and predicate terms EXACTLY as in the source language.

STRICT RULES:

1. FORMAT
Each sentence must follow:
[Quantifier] [native_subject] are [native_predicate]

Allowed quantifiers:
All / No / Some / Some...not

Never output:
"All X are not Y"

2. DO NOT MODIFY LOGIC
- Do NOT swap subject and predicate.
- Do NOT replace terms.
- Do NOT merge synonyms.
- Do NOT normalize or repair the argument.
- Do NOT balance term distribution.
- Do NOT split polysemous words.
- Do NOT introduce [s] or [g] tags.

Extract the argument exactly as written.

3. GOOGLE TRANSLATE CHECK (Fidelity Only)
Mentally compare with Google Translate ONLY to:
- Verify quantifier accuracy
- Verify negation scope
- Verify copula meaning

4. TERM HANDLING
- Copy subject and predicate verbatim.
- Singular/plural variants count as the same term.
- Descriptive phrases remain descriptive phrases.
- Identity statements (X are X) remain unchanged.

5. TERM COUNT CHECK (Diagnostic Only)
After extraction:
- Count distinct terms.
- A classical syllogism should have exactly 3 distinct terms.
- If not, DO NOT repair.
- Simply report the count in reasoning.

6. SENTENCE SELECTION
If more than 3 sentences are present:
- Select the two premises and the conclusion that form the main argument.
- Ignore rhetorical commentary.

REASONING FORMAT:
1. Identify selected sentences (P1, P2, C)
2. Extract subject and predicate
3. Count distinct terms (no fixing)
4. Output final structured form

OUTPUT (JSON only, no markdown):
{
  "detected_language": "<lang>",
  "reasoning": "<selection + extraction + term_count>",
  "english": "P1. P2. Therefore, C."
}
```

Figure 7: EPN prompt with google translated sentence for Subtask 4 for extract and filter relevant premise and conclusion