Lakksh at SemEval-2026 Task 11(1 & 2): Neuro-Symbolic Decomposition to Mitigate Content Bias in Syllogistic Reasoning

Lakksh Sharma
Thapar University, Patiala, India
lakksh.sharma@gmail.com

Krish Sharma
Thapar University, Patiala, India
ksharma8_be23@thapar.edu

Jatin Bedi
Thapar University, Patiala, India
jatin.bedi@thapar.edu

Abstract

Syllogistic reasoning is the ability to distinguish logical validity from semantic plausibility — a setting in which LLMs succumb to frequent content bias by conflating the two. The result is a characteristic failure to recognize logically valid arguments with highly implausible conclusions and logically invalid but semantically plausible arguments. This paper introduces a neuro-symbolic system that avoids this behavior by design: neural structure extraction is strictly separated from symbolic validity checking. A T5-Small parser is trained only on synthetic nonsense-symbol syllogisms, ensuring that the structural parse is learned in the absence of real-world semantics. Validity checking is performed by a deterministic symbolic kernel operating on extracted logical form alone, ensuring that semantic content cannot influence the final call. In binary validity classification, the system achieves 97.38% accuracy with a Total Content Effect of 3.10; in the retrieval setting, it achieves 82.11% accuracy with 99.47% F1 on premise identification. Ablation experiments show that formal theorem proving via NL-to-Z3 translation actually increases content bias due to leakage in intermediate representations. The results recommend architectural separation as a promising content-robustness strategy for syllogistic reasoning. [1]

1 Introduction

Syllogistic reasoning is the task of determining whether a conclusion is valid given a set of premises, regardless of the real-world plausibility of the statements being combined. While syllogistic validity is fully characterized in classical logic, recent work has shown that large language models systematically conflate logical validity with semantic plausibility, leading to systematic content effects in reasoning judgments (Dasgupta et al., 2022). Loosely speaking, logically valid but semantically implausible arguments are often denied, while plausible but logically invalid arguments are often accepted. This issue has been noted across syllogistic benchmarks and evaluation settings, including English-language syllogism benchmarks (Ozeki et al., 2024; Bertolazzi et al., 2024).

This paper focuses on English syllogistic reasoning under explicit content-bias evaluation, such that models must predict logical validity while minimizing sensitivity to semantic plausibility. Prior work has proposed to mitigate content effects via a variety of prompting strategies, explanation refinement, or interventions at the level of activations (Valentino et al., 2025; Lyu et al., 2023). While these approaches can mitigate bias in certain settings, they keep semantic content and logical decision making coupled within the same representational pathway. The authors take the perspective that content bias is fundamentally an architectural issue: as long as semantic content is accessible by the validity decision making component, bias can always resurface.

Hence, the paper presents a neuro-symbolic system that explicitly separates the structure extraction process from the step of validity verification. The neural component is used only to map the natural language syllogism to an abstract logical form and the validity is checked by a deterministic symbolic kernel that operates only on this abstraction. To eliminate leakage of semantic information during structure extraction, the neural parser is trained only with synthetic syllogisms written in nonsense symbols. As such, it never learns any association between any real-world entity and its validity. The authors observe that such a separation of architecture can greatly reduce content bias, achieving high accuracy with low Total Content Effect on the English syllogistic validity classification task. Ablation experiments further show that even formally grounded approaches such as NL-to-symbolic the-

[1] https://github.com/lkksharma/Semeval-11-12

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 115–120

orem proving can increase content bias if the intermediate representations reintroduce semantic information.

2 Background

2.1 Task Setup

Syllogistic validity classification requires determining whether a conclusion logically follows from two natural-language premises. Task 11 (Valentino et al., 2026) evaluates systems under content bias—the tendency for predictions to rely on semantic plausibility rather than formal validity. This paper addresses the English subtasks: Subtask 1 (binary validity) and Subtask 2 (premise retrieval + validity).[2]

Subtask 1 (Binary Validity). Given a syllogism, predict VALID or INVALID.
Example: "All cats are mammals. All mammals are animals. Therefore, all cats are animals." → VALID

Subtask 2 (Premise Retrieval + Validity). Given a paragraph with distractors, identify the two relevant premise indices and predict validity.
Example: [1] "The sky is blue." [2] "All mammals are warm-blooded." [3] "All dogs are mammals." [4] "Therefore, all dogs are warm-blooded." → premises=[2,3], VALID

The evaluation metric rewards both high accuracy and low sensitivity to plausibility. Plausibility and validity are statistically independent in the English tracks, enabling clean measurement of content effects.

2.2 Related Work

Content effects in LLMs. Language models exhibit belief-bias on syllogistic tasks, treating plausibility as a proxy for validity (Dasgupta et al., 2022). NeuBAROCO provides evidence of such biases (Ozeki et al., 2024), while broader studies characterize error modes in quantifier scope and negation handling (Bertolazzi et al., 2024; Eisape et al., 2024).

Mitigation strategies. Activation steering modulates content bias at inference time (Valentino et al., 2025). Quasi-symbolic methods improve chain-of-thought reliability via structured intermediate forms (Ranaldi et al., 2025). Symbolic provers can verify natural language explanations, though this requires robust NL-to-logic translation (Quan et al., 2024; Xu et al., 2024). Mechanistic interpretability suggests reasoning components may be localized in transformers, motivating explicit structure-content separation (Kim et al., 2025).

The present approach. This work imposes a stricter constraint: the validity-deciding component never accesses semantic content. The system decomposes reasoning into structure extraction (NL → Mood/Figure) and deterministic verification, guaranteeing content-invariance architecturally rather than via behavioral interventions.

3 System Overview

The information barrier proposed is strict: **the component that makes the logical decision never accesses the semantic content**. The authors achieve this by decomposing syllogistic reasoning into two separate steps: (i) neural structure extraction, which translates natural language into abstract logical form, and (ii) symbolic verification, which ascertains validity based on that abstraction alone. For Subtask 2, a retrieval phase initially selects the pertinent premises from the disorderly paragraphs.

3.1 Architecture

The pipeline consists of four stages, each with a clear interface:

1. **Retrieval (Subtask 2 only):** An LLM-based classifier identifies the two sentences serving as premises. The model is prompted to track term occurrences—identifying sentences containing the major term (conclusion predicate) and the minor term (conclusion subject). Failures typically cluster around complex pronominal coreference.

2. **Structure Extraction:** The transformation of natural-language syllogisms to a canonical (Mood, Figure) representation. The three types of propositions (A/E/I/O) are encoded by Mood, and the arrangement of terms is encoded by Figure. This reduces the classification problem to 1) quantifier identification and 2) predicate-argument structure extraction.

3. **Ensemble Fusion:** amnesiac T5 and Llama-3 individually extract the (Mood, Figure) tuple, and their outputs are OR-fused: if either extractor outputs a valid form, the system out-

[2] https://github.com/neuro-symbolic-ai/semeval_2026_task_11

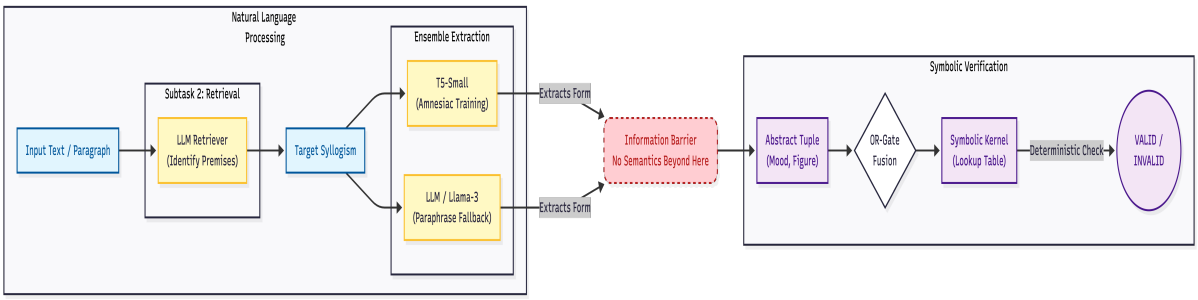

Figure 1: **System Architecture.** The pipeline is strictly divided by an information barrier into two distinct phases. **Left (Semantic Phase):** The system processes natural language inputs using an LLM-based retriever (for Subtask 2) and an ensemble of structure extractors, including an Amnesiac T5 trained on nonsense symbols and a fallback LLM. **Right (Symbolic Phase):** The extracted logical forms—represented solely as (Mood, Figure) tuples—are verified by a deterministic Symbolic Kernel. This architectural separation ensures that the final validity decision is mathematically incapable of accessing semantic content.

puts Valid. This capitalizes on complementary failure modes, T5 on canonical forms and Llama-3 on paraphrases, to recover cases where a single extractor is fooled.

4. **Symbolic Kernel:** A deterministic look-up table encoding the 15 valid Aristotelian forms. This component has no trainable parameters and receives only the abstract (Mood, Figure) tuple, not the original text.

The cardinal rule of the design is that natural language is only processed during extraction. A content-blind symbolic component makes the final validity decision, ensuring semantic plausibility cannot influence the logical judgment.

3.2 Amnesiac Training

The main technical challenge is learning a structure extractor that acquires syntactic but not semantic associations. Standard fine-tuning on real syllogisms would encode correlations between entity types and validity outcomes—precisely the content bias the authors aim to avoid.

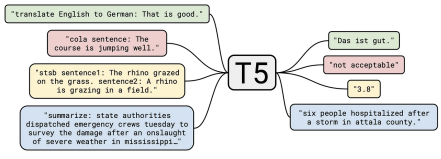

Figure 2: T5 reformulates NLP tasks as text-to-text generation with a shared encoder–decoder architecture (Raffel et al., 2020).

The authors propose **Amnesiac Training**: training a T5-Small model solely on synthetic syllogisms constructed with meaningless placeholder symbols (ALPHA, BETA, GAMMA, etc.).

Training Instance Example:
Input: All ALPHA are BETA. No BETA is a GAMMA. Therefore, no ALPHA is a GAMMA.
Output: Mood: AEE, Figure: 1

Dataset Construction. The authors generate instances for all combinations of 256 (Mood, Figure) tuples by: (1) stochastically sampling 3 placeholders without replacement for subject, middle, and predicate terms; and (2) applying paraphrase templates to each proposition. Linguistic variation is encoded via templates (e.g., "All X are Y" vs. "Every X is a Y" vs. "Each X is a Y"). Approximately 50 templates per proposition type are produced, resulting in 200k training examples.

Why This Works. Since placeholders have no semantic meaning, the model cannot learn associations between validity and specific concepts (e.g., that valid syllogisms are associated with "scientists" and invalid ones with "unicorns"). It learns only: (1) quantifier identification, (2) polarity detection, and (3) term position patterns. The training signal is strictly structural.

Configuration. T5-Small (60M parameters) is used with the AdamW optimizer, a learning rate of 10^{-4}, 1k-step warmup, batch size 32, for 10 epochs. Inference uses greedy decoding parsed via regex.

Limitations. The amnesiac parser struggles with structures not seen in templates, such as complex negations ("It is not the case that...") or passive voice inversion. These edge cases are handled by the ensemble fallback.

3.3 Ensemble Fusion

The extractors are characterized by complementary error patterns:

Extraction Tool	Strength	Weakness
T5 (Amnesiac)	Canonical forms	Paraphrases
LLM (Llama-3)	Paraphrases	Content bias

Table 1: Complementary strengths of the ensemble components.

OR-Fusion Strategy. The system combines extractors at the decision level: if *any* extractor predicts a valid syllogistic form, the system outputs VALID. This asymmetric synthesis prioritizes recall for valid syllogisms.

Rationale. Content bias is mostly in the form of false negatives, that is, the models are not accepting valid-but-implausible syllogisms because the conclusion appears to be semantically incorrect. With the union of valid predictions, the system extracts those examples where one of the extractors identified the validity correctly, although the other was misled by content.

3.4 Symbolic Kernel

Validity checking is a pure look-up operation:

The symbolic kernel encodes the 15 valid Aristotelian forms—AAA-1, EAE-1, AII-1, EIO-1, EAE-2, AEE-2, EIO-2, AOO-2, AII-3, IAI-3, EIO-3, OAO-3, AEE-4, IAI-4, EIO-4—as a lookup set. Validity checking reduces to set membership: `is_valid(mood, figure) = (mood, figure)` $\in$ `VALID`.

This component is: (1) **Parameter-free**; (2) **Content-blind** (inputs are abstract tuples, not text); (3) **Deterministic**; and (4) **Interpretable** (decisions map directly to Aristotelian logic). This ensures that regardless of upstream extraction errors, the ultimate validity verdict rests on logical form alone.

Negative Result: Z3 Theorem Proving. The authors also explored translating syllogisms to Z3 constraints using LLM-generated Python code. On the contrary, the Total Content Effect (TCE) increased with the Z3 approach.

Root Cause. The NL $\rightarrow$ Z3 translation step itself is content-biased: (1) variable naming leaks entity semantics; (2) constraints formed from plausible premises are more likely to be well-formed; and (3) the LLM's priors regarding set cardinality influence quantifier translation. The solver functions correctly, but operates on biased representations.

Lesson. Formal verification alone is insufficient; the translation layer must also be content-agnostic. The amnesiac extraction approach addresses this specific vulnerability.

Design Choice	Bias Mitigation
Symbolic bottleneck	Kernel sees no text
Amnesiac training	No semantic correlations
OR-ensemble fusion	Recovers false negatives
Zero-param kernel	No learnable bias

Table 2: Architectural invariants targeting content bias.

These are architectural constraints, not prompt heuristics.

4 Experimental Setup

Table 3 summarizes the data and model configuration.

Component	Configuration
Training data	200K synthetic (nonsense-symbol)
Evaluation data	960 English syllogisms
T5-Small	10 ep, bs 32, lr 10^{-4}, 1K warmup
Retrieval LLM	Gemini-2.0-Flash, temp 0.2
Extraction LLM	Llama-3 (fallback)
Symbolic Kernel	Deterministic lookup (0 params)

Table 3: Experimental configuration.

Data and Preprocessing. The 960 official English syllogisms are used only for evaluation; the amnesiac T5 parser trains exclusively on 200K synthetic syllogisms with placeholder symbols, ensuring zero exposure to real entities. Premises and conclusion are concatenated canonically and tokenized (T5Tokenizer, max length 128) without lowercasing or lemmatization to preserve quantifier semantics. For Subtask 2, retrieved premises are concatenated with the conclusion before extraction. No external data is used.

Evaluation. Subtask 1 reports Accuracy (ACC), Total Content Effect (TCE), and Combined score:

$$\text{Combined} = \frac{\text{ACC}}{1 + \ln(1 + \text{TCE})} \tag{1}$$

This metric rewards accuracy while penalizing content sensitivity. Subtask 2 additionally reports F_1 for premise identification. A perfect system achieves ACC = 100% and TCE = 0.

5 Results

5.1 Main Results

System	ACC	TCE	Comb.
Subtask 1 (Binary Validity)			
Llama-3 only	90.05	7.29	28.91
Amnesiac T5 only	70.16	33.33	15.47
Ensemble + Kernel	**97.38**	**3.10**	**40.38**
Subtask 2 (Retrieval + Validity)			
Ensemble + Kernel	82.11	9.89	26.80

Table 4: Main results on Codabench test set. Subtask 2 F_1: 99.47%.

Subtask 1 (Binary Validity). The final system (LLM + amnesiac T5 with OR fusion and symbolic kernel) achieves 97.38% accuracy with TCE = 3.10, yielding a Combined score of 40.38. This substantially improves over the LLM-only baseline (ACC 90.05, TCE 7.29), indicating both higher accuracy and reduced sensitivity to semantic plausibility.

Subtask 2 (Premise Retrieval + Validity). The retrieval-first pipeline achieves 82.11% accuracy on validity prediction with 99.47% F_1 on premise identification (Combined 26.80). These results demonstrate reliable premise selection while maintaining content-robust validity judgments. A single baseline Llama-3 for Subtask 2 was not tested since the retrieval stage (Gemini) is the same for all configurations and the extraction-level ablation is already contained in the Subtask 1 comparison.

5.2 Ablations

Z3 Usage	ACC	TCE	Comb.
0%	82.11	9.89	26.80
50%	83.33	17.79	22.97
100%	76.32	25.00	18.91

Table 5: Z3 ablation (Subtask 2): percentage of instances where the symbolic kernel was replaced by NL→Z3 verification. Higher Z3 usage amplifies TCE.

Ensemble Fusion. Combining the LLM and amnesiac T5 parser yields +7.33% accuracy and −4.19 TCE compared to the LLM alone (Table 4). Syntactic extraction from the amnesiac parser complements the LLM's paraphrase robustness.

Amnesiac T5 Only. The T5-only configuration achieves lower accuracy (70.16%) with high TCE (33.33). Parsing failures default to INVALID, creating systematic bias unrelated to content. This motivates ensemble fusion rather than standalone use.

NL→Z3 Theorem Proving. Z3-based verification increases accuracy marginally but consistently amplifies TCE, worsening Combined scores as Z3 usage increases (Table 5). This confirms that raw accuracy gains can be offset by content bias when intermediate representations leak semantic information.

5.3 Error Analysis

Manual analysis of 50 randomly sampled errors reveals four failure modes:

Parsing Errors. Complex negation and scope handling cause misclassification. Double negation (e.g., "It is not the case that every philosopher is a mammal") leads to incorrect mapping between universal and particular forms.

Figure Misidentification. Incorrect syllogistic figure assignment occurs under non-canonical word order or embedded clauses.

Residual Content Bias. The LLM occasionally rejects valid-but-implausible syllogisms (e.g., "All scientists are artists"). In such cases, the amnesiac T5 parser often predicts the correct structure, allowing OR fusion to recover validity. Ensemble fusion corrects approximately 60% of content-driven errors observed in LLM-only predictions.

Performance Gap. The reduction in accuracy (82.11%) of Subtask 2 despite high retrieval F_1 (99.47%) is due to premise ordering errors. Although the retriever accurately isolates the relevant sentences, it sometimes fails to canonically order the major and minor premises, causing the structure extractor to misidentify the syllogistic Figure.

6 Conclusion

This paper presented a neuro-symbolic system that mitigates content bias through architectural separation: neural extraction maps syllogisms to abstract (Mood, Figure) tuples, while a deterministic symbolic kernel verifies validity without accessing semantic content. Training on synthetic nonsense syllogisms (Amnesiac Training) prevents entity–validity associations. The system ranked 24th in Subtask 1 and 12th in Subtask 2. A key negative result is that NL→Z3 theorem proving amplifies content bias despite theoretical soundness—*where* symbolic reasoning is introduced matters as much as *whether* it is used. Future work includes permutation-agnostic premise validation and generating interpretable explanations from symbolic derivations.

References

Luca 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 (EMNLP)*.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Gyuwan Kim, Marco Valentino, and André Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Qing Lyu, Shreyas 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 (AACL)*.

Kento Ozeki, Ryo Ando, Terufumi Morishita, Hitomi 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*.

Xiao Quan, Marco Valentino, Louise Dennis, and André 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 (EMNLP)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Leonardo Ranaldi, Marco Valentino, and André Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marco Valentino, Gyuwan Kim, Dhara Dalal, Zhibin Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *arXiv preprint arXiv:2505.12189*.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026. Semeval-2026 task 11: Disentangling content and formal reasoning in large language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Jiarui Xu, Hao Fei, Li Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

Appendix

A.1 Synthetic Templates. Proposition templates for A/E/I/O types:
A: "All X are Y", "Every X is a Y"
E: "No X is a Y", "Not a single X is Y"
I: "Some X are Y", "At least one X is Y"
O: "Some X are not Y", "Not all X are Y"

A.2 File Descriptions. *Subtask 1 (Binary Validity):*
`symbolic_syllogism_engine.py`: Symbolic kernel + T5/LLM/rule-based parsers
`fuse_gemini_t5.py`: OR fusion of LLM + T5 predictions
Training:
`train_synthetic_parser.py`: Amnesiac T5 training on synthetic data
`content_anonymizer.py`: Entity anonymization utilities
Subtask 2 (Retrieval + Validity):
`subtask2_engine.py`: Retrieval + symbolic validation pipeline
`z3_solver.py`: Z3 theorem prover integration