YNU-NLP at SemEval-2026 Task 11: A Neuro-Symbolic Approach with Reflexion Mechanism Disentangling Content and Formal Reasoning in Language Models

Yu Wang, You Zhang*, Hao Zhang, and Dan Xu
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: WangY@stu.ynu.edu.cn, yzhang0202@ynu.edu.cn

Abstract

This paper describes our systems for SemEval-2026 Task 11, Disentangling Content and Formal Reasoning in Language Models. We participated in all four subtasks, addressing the *Content Effect*—a phenomenon where models rely on real-world plausibility rather than logical validity. Existing methods, such as standard Chain-of-Thought (CoT) prompting or single-task Supervised Fine-Tuning (SFT), often struggle to completely decouple content from reasoning due to the inherent probabilistic biases in pre-trained models. To address these limitations, a hybrid neuro-symbolic framework based on the Qwen2.5-14B architecture is proposed, integrating multi-task instruction tuning with a robust neuro-symbolic pipeline. The principal innovation lies in the deployment of a *Reflexion* mechanism coupled with formal verification: natural language arguments are parsed into First-Order Logic (FOL) and subsequently verified by the Z3 Theorem Prover. Parsing anomalies are automatically rectified through an iterative self-correction module. The proposed system ranked 1st in Subtasks 1 & 2, 2nd in Subtask 4, and 9th in Subtask 3, validating its ability to decouple logical validity from content plausibility. To strictly ensure replicability, the dataset and codebase are available at https://github.com/DALUOLUODE/Semeval_2026_task11.

1 Introduction

While Large Language Models (LLMs) demonstrate remarkable proficiency in text generation (Brown et al., 2020), their reasoning capabilities are often predicated on probabilistic patterns and pre-trained world knowledge rather than rigorous logical rules. Consequently, a fundamental misalignment often occurs between formal validity and semantic plausibility (Dasgupta et al., 2024; Eisape et al., 2024). To systematically investigate and mitigate this phenomenon, SemEval-2026 Task 11 (Valentino et al., 2026; Ghosh et al., 2026) was introduced, challenging systems to disentangle formal reasoning from content biases across multilingual settings. This task fundamentally builds upon recent mechanistic interpretations of syllogistic inference (Kim et al., 2025) and efforts to mitigate content effects through fine-grained activation steering (Valentino et al., 2025).

To mitigate the reliance on pre-trained priors, strategies such as CoT (Wei et al., 2022; Kojima et al., 2022; Ranaldi et al., 2025) and SFT have been proposed. However, purely neural methods often fail to completely decouple reasoning from the probabilistic associations of world knowledge. Furthermore, existing neuro-symbolic (Pan et al., 2023; Olausson et al., 2023; Quan et al., 2024; Xu et al., 2024) research is predominantly English-centric (Wu et al., 2023), leaving a significant gap in understanding how content biases manifest in complex multilingual and low-resource environments.

To address these limitations, a hybrid Neuro-Symbolic system based on the Qwen2.5-14B architecture is proposed. We employ a dual-mode approach where the model functions both as a semantic parser—translating arguments into FOL for verification by the Z3 Theorem Prover—and as a direct reasoning solver.

The contributions of this work are summarized as follows:

- A high-quality, augmented dataset was constructed. The original dataset was expanded fourfold through structural enhancement, noise injection for retrieval subtasks, and multilingual translation to simulate low-resource settings.
- A robust *Try-Fix-Verify* architecture is introduced. This pipeline incorporates a novel "Re-

*Corresponding author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 919–926

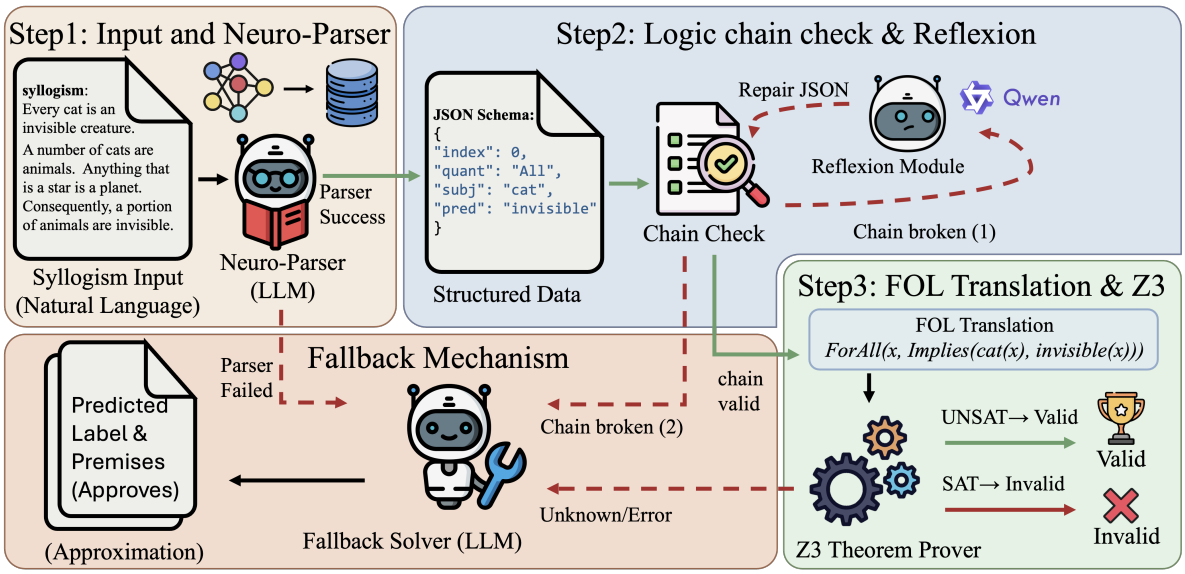

Figure 1: The architecture of the proposed Neuro-Symbolic system.

flexion" mechanism to automatically repair term mismatches (e.g., "tall" vs. "very tall") during parsing, thereby reducing failures in symbolic execution.

- A Multi-task SFT strategy is implemented. By unifying structure extraction and logical reasoning within a single model, the system's ability to interpret complex logical structures is enhanced.

Through this framework, our system achieved highly competitive results: ranking 1st in Subtasks 1 & 2, 2nd in Subtask 4, and 9th in Subtask 3. Analyses demonstrate that the Z3 prover drastically reduces the Total Content Effect (TCE), effectively disentangling logical validity from semantic plausibility.

The subsequent sections detail the neuro-symbolic pipeline (Section 2), experimental setup (Section 3), results and analysis (Section 4), and conclusions (Section 5).

2 System Overview

As illustrated in Figure 1, a hybrid framework is proposed combining neural semantic parsing with symbolic verification to disentangle formal reasoning from content plausibility. The architecture integrates a Neuro-Parser for extracting logical schemas, a Reflexion Mechanism for structural repair, and a Symbolic Verifier based on FOL. A Neural Solver serves as a fallback to ensure robustness.

2.1 Data Construction & Augmentation

To surmount the inherent constraints of data scarcity and to explicitly optimize the model for complex multilingual and retrieval-augmented reasoning scenarios, the original training set was systematically expanded fourfold to 3,836 instances. This was achieved through a rigorous pipeline involving JSON-based structural enhancement (via Gemini 3 Pro), distractor noise injection (1-5 irrelevant premises), and multilingual expansion across 11 target languages (via Qwen-MT-Plus). Detailed procedures and data examples are provided in Appendix A.

2.2 Multi-task Instruction Tuning

A single unified model was trained to perform both structural parsing and direct reasoning using Qwen2.5-14B-Instruct as the base model. To reduce memory overhead, the model was loaded with 4-bit quantization and fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021).

Task Formulation. The training data was formatted into two distinct instruction-based tasks:

1. Parser Task: Guided by `PARSER_INSTRUCTION`, the model is trained to translate natural language input into a strict JSON schema defining the logical structure.

2. Solver Task: Guided by `SOLVER_INSTRUCTION`, the model is trained to directly predict the validity label and

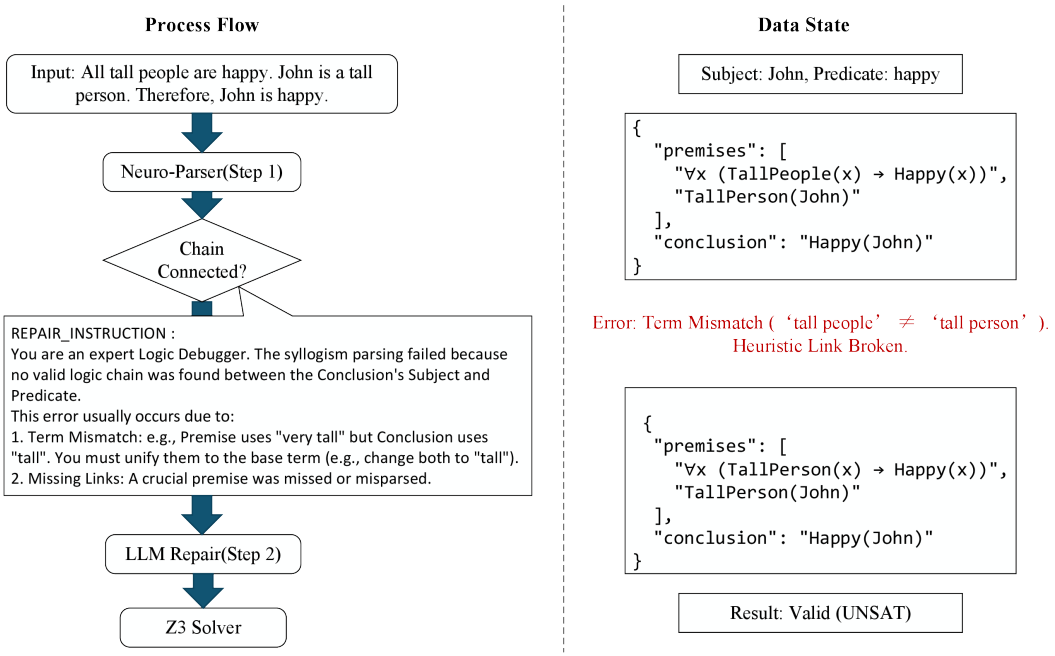

Figure 2: Reflexion mechanism case study.

identify relevant premises based on internal reasoning.

Training Strategy. To prevent the model from overfitting to the instruction templates, a response-only masking strategy was utilized. This technique ensures that the loss is calculated exclusively on the generated output (the JSON response), masking the instruction and input text.

2.3 Neuro-Symbolic Pipeline with Reflexion

As illustrated in Figure 2, the inference process follows a robust *Try-Fix-Verify* pipeline. Initially, the Neuro-Parser generates a JSON representation of the syllogism structured via `PARSER_INSTRUCTION`. A heuristic algorithm is subsequently deployed to validate the logical chain. Specifically, this module meticulously identifies the subject (S), predicate (P), and the middle term (M) across the parsed premises, attempting to reconstruct the deductive linkage. If the S-M-P chain is identified as broken due to term inconsistencies (e.g., "tall" versus "very tall"), the Reflexion Mechanism (Shinn et al., 2023; Madaan et al., 2023) is immediately triggered. Instead of failing immediately, the system prompts the model with `REPAIR_INSTRUCTION` to unify inconsistent terms and regenerate the structure, thereby resolving parsing errors that would otherwise block symbolic execution.

Upon successful extraction, the Symbolic Verifier translates the logical structure into FOL for the Z3 Theorem Prover. Validity is determined via proof by contradiction: the system asserts the premises and the negation of the conclusion; if Z3 (De Moura and Bjørner, 2008) returns UNSAT, the argument is logically valid. In cases where symbolic verification fails (e.g., due to unresolvable parsing errors or timeouts), a Fallback Strategy is employed, reverting to the neural solver (`SOLVER_INSTRUCTION`) to ensure a robust end-to-end prediction.

3 Experimental Settings

3.1 Evaluation Metrics

We strictly adopt the official evaluation metrics defined for SemEval-2026 Task 11 (Valentino et al., 2026), including Accuracy (ACC), Premise F1-score, Total Content Effect (TCE), and the unified primary Ranking Score.

3.2 Baselines

We compare our method against Zero-shot Prompting (Kojima et al., 2022), CoT (Wei et al., 2022), and a standard SFT baseline that predicts validity directly without the symbolic pipeline. Unless otherwise stated, all trainable baselines were trained on the same 3,836-instance augmented training set described in Section 2.1, so that the comparison isolates the contribution of the neuro-symbolic in-

ference pipeline rather than differences in training data scale.

3.3 Implementation Details

We fine-tuned Qwen2.5-14B-Instruct (4-bit quantized) using LoRA ($r = 16$, $\alpha = 32$) via the Unsloth library[1] for efficient training. Training used a batch size of 8, learning rate 2×10^{-4}, and 5 epochs with a response-only mask. Inference temperature was set to 0.01 to ensure deterministic JSON generation. Experiments were conducted on a single NVIDIA GeForce RTX 3090 (24GB) GPU.

4 Experimental Results

4.1 Main Results on the Official Test Set

As summarized in Tables 1 and 2, the proposed Neuro-Symbolic architecture achieved the highest ranking scores. Crucially, substantial gains in robustness are demonstrated: in Subtask 4, the TCE dropped significantly from 14.65 (Standard SFT) to 1.26 (Ours). It is empirically demonstrated that formally verifying the logic via the Z3 solver effectively disentangles reasoning from content plausibility.

4.2 Ablation Study

To verify the contribution of key components—namely the Reflexion Mechanism and the Symbolic Solver—an ablation study was conducted on the Subtask 4 Official Test Set.

Effect of Reflexion Mechanism. As shown in Table 3, enabling the Reflexion mechanism resulted in a significant improvement in the primary Ranking Score and Premise F1. By successfully resolving term inconsistencies via the *Try-Fix* loop, the system can process more instances through the rigorous symbolic solver. In the w/o Reflexion setting, these instances would fail the heuristic check and trigger the fallback strategy, leading to a higher TCE and lower overall robustness.

Effect of Symbolic Solver (Z3). To verify the structural necessity of the Symbolic Solver module, we performed an ablation by removing the Z3 verification component entirely (effectively reducing the pipeline to a pure neural model, equivalent to the **Standard SFT** baseline). For this experiment, the Reflexion mechanism was deactivated in both settings to ensure a direct comparison of the reasoning engines' inherent capabilities.

As presented in Table 3, the results indicate a fundamental difference in how the models handle content bias. The Pure SFT model, while achieving competitive Accuracy, exhibits a significantly higher TCE (14.65 vs. 3.04). This suggests that the pure neural model relies more heavily on content plausibility shortcuts. In contrast, the Neuro-Symbolic approach (even without Reflexion) achieves a much lower TCE, confirming that translating natural language into First-Order Logic effectively shields the reasoning process from content interference. Consequently, the Neuro-Symbolic approach yields a superior Ranking Score.

4.3 Analysis

Cross-Lingual Robustness and Generalization. Instead of relying on language-specific logical structures, our system employs a translation-based parsing strategy where the Neuro-Parser normalizes input from diverse languages into a unified English-based JSON schema. As observed in Table 2, this approach demonstrates remarkable stability. When transitioning from the monolingual English setting (Subtask 2) to the complex multilingual setting (Subtask 4), the Ranking Score actually increased from 46.23 to 49.51, despite the increased task difficulty. While the Accuracy showed a moderate decline (from 98.96% to 90.10%), the system maintained a high level of robustness. This improvement in the final score is driven by a superior TCE reduction in the multilingual setting (1.26 vs. 2.13), indicating that the Neuro-Symbolic pipeline effectively decouples reasoning from linguistic surface forms (Clark et al., 2020; Bertolazzi et al., 2024). By separating the linguistic representation (handled by the Parser) from the reasoning engine (handled by Z3), the system avoids the common pitfall where low-resource languages suffer from degraded reasoning capabilities due to insufficient pre-training data. However, we acknowledge a trade-off: while our system excelled in noisy environments (Subtask 4), its lower ranking in the clean multilingual subtask (Subtask 3) suggests that the current pipeline may be less effective when symbolic verification is applied to already fluent multilingual inputs. One plausible explanation is that the strict verification stage can reject some cases that a purely neural model would answer correctly. A more fine-grained error analysis is needed to determine whether the main source of errors comes from parsing, translation normalization, or

[1] https://github.com/unslothai/unsloth

Model	Subtask 1 (English)			Subtask 3 (Multilingual)		
	ACC	TCE	Score	ACC	TCE	Score
Zero-shot (`Qwen2.5`)	61.26	41.67	12.89	63.02	42.71	13.19
CoT	85.86	20.17	21.19	80.21	12.50	22.26
Standard SFT	98.95	1.04	57.74	93.75	8.33	28.99
Ours (Neuro-Symbolic)	100.00	0.00	100.00	94.79	6.25	31.80

Table 1: Main results on the official test set for binary validity classification (Subtasks 1 & 3). Note that the Ranking Score is based on ACC and TCE.

Model	Subtask 2 (English Retrieval)				Subtask 4 (Multilingual Retrieval)			
	ACC	TCE	F1	Score	ACC	TCE	F1	Score
Zero-shot (`Qwen2.5`)	58.33	42.71	12.32	7.39	54.69	42.71	13.02	7.09
CoT	81.77	12.09	82.16	22.95	74.48	25.60	72.99	17.22
Standard SFT	92.19	11.46	81.77	24.69	82.81	14.65	65.62	19.79
Ours (Neuro-Symbolic)	98.96	2.13	98.96	46.23	90.10	1.26	89.58	49.51

Table 2: Main results on the official test set for validity classification with premise retrieval (Subtasks 2 & 4). Note that the Ranking Score considers Premise F1.

Model	Subtask 4 (Multilingual Retrieval)			
	ACC	TCE	F1	Score
Ours (Full)	**90.10**	**1.26**	**89.58**	**49.51**
w/o Reflexion	88.54	3.04	84.38	36.09
w/o Symbolic	82.81	14.65	65.62	19.79

Table 3: Comprehensive ablation results on the official test set of Subtask 4.

conservative symbolic filtering.

Retrieval Capability under Noise. In Subtasks 2 and 4, the critical challenge is distinguishing between logically necessary premises and irrelevant noise (distractors). The results in Table 2 reveal a significant disparity in the Premise F1-score between the Standard SFT baseline (65.62) and our Neuro-Symbolic approach (89.58) in the multilingual setting. This performance gap highlights the structural advantage of the symbolic pipeline. The Standard SFT model operates end-to-end and often hallucinates relevance based on semantic similarity—mistaking noisy premises that are topically related to the conclusion for logically relevant ones. In contrast, our pipeline utilizes the Z3 Theorem Prover to construct a formal proof by contradiction. The heuristic linkage and symbolic verification processes inherently act as a rigorous filter: premises that do not contribute to the deductive chain connecting the Subject and Predicate are automatically excluded. Consequently, the set of relevant premises identified by our system is derived from logical necessity rather than statistical probability, leading to superior precision even in high-noise environments.

5 Conclusion

In this paper, we present a system for SemEval-2026 Task 11 that disentangles formal validity from content plausibility in multilingual syllogistic reasoning. By integrating multi-task SFT with a neuro-symbolic pipeline, we combine the semantic flexibility of Large Language Models with the rigor of theorem proving.

Experimental results show that our system achieves competitive accuracy while substantially reducing the TCE, ranking 1st in Subtasks 1 & 2, 2nd in Subtask 4, and 9th in Subtask 3. These results confirm that formal verification is an effective way to mitigate belief bias in logical tasks (Ozeki et al., 2024). Our framework also identifies relevant premises robustly under noise and remains stable across diverse linguistic settings.

For future work, we plan to scale the model from 14B to 72B to improve multilingual parsing, compare the framework with other neuro-symbolic systems such as LINC and Logic-LM, and extend the symbolic engine to support richer logical forms such as modal logic.

Acknowledgments

This work was supported by the Scientific Research Fund of Yunnan Provincial Education Department under Grant No. 2026J0006 and by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

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 (EMNLP 2024)*, pages 13882–13905.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pages 3882–3890.

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. *arXiv preprint arXiv:2207.07051*.

Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: An efficient SMT solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2008)*, pages 337–340.

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 (NAACL 2024)*, pages 8425–8444.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*.

Geonhee Kim, Marco Valentino, editor = "Che Wanxiang Freitas, Andre", Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*, pages 22199–22213.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: Iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, pages 46534–46594.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 5153–5176.

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.

Liangming Pan, Alon Albalak, Xinyi Wang, and William 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.

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 (EMNLP 2024)*, pages 2933–2958.

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 (ACL 2025)*, pages 17222–17240.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R. Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *arXiv preprint arXiv: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)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*, pages 24824–24837.

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.

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 (ACL 2024)*, pages 13326–13365.

A Data Construction & Augmentation Details

As outlined in Section 2.1, the official training dataset was expanded fourfold to optimize the model for complex multilingual and retrieval-augmented reasoning scenarios. This appendix provides the detailed procedures and concrete data examples for each step of the pipeline.

A.1 Detailed Pipeline Procedures

Step 1: Structural Enhancement. To establish a high-quality baseline schema, the original 959 unstructured English syllogisms were transformed into a rigorous JSON format. This initial extraction—explicitly delineating premises, conclusions, quantifiers, subjects, and predicates—was facilitated via Gemini 3 Pro and subsequently subjected to manual verification. This ensures the foundational logic chain is flawlessly structured before any noise or translation is applied.

Step 2: Noise Injection. A "Rich Noise Pool" of unique source premises was extracted from the structurally enhanced data. For retrieval subtasks, a heuristic script randomly selected 1 to 5 irrelevant premises (distractors) from the pool and injected them into each instance. The original indices of the logically necessary premises were recorded as the `relevant_premises` array, which serves as the ground truth for Subtasks 2 and 4.

Step 3: Multilingual Expansion. To support cross-lingual subtasks and simulate low-resource settings, the English dataset was translated into the remaining 11 official target languages (German, Spanish, French, Italian, Dutch, Portuguese, Russian, Chinese, Swahili, Bengali, and Telugu). The Qwen-MT-Plus API was utilized with a round-robin strategy (`lang_idx = idx % 11`) to distribute the target languages evenly across the generated instances.

Final Dataset Composition. The final dataset comprises four balanced subsets (959 instances each) perfectly aligned with the SemEval subtasks: (1) English Clean, (2) English Noisy, (3) Multilingual Clean, and (4) Multilingual Noisy. This yields a total of 3,836 diverse training instances.

A.2 Data Transformation Examples

The following examples illustrate the step-by-step transformation of a single training instance through our augmentation pipeline.

1. Original Official Data (English Clean): The original data provides only the raw text, validity, and plausibility labels.

```
{
  "id": "0",
   "syllogism": "All cars are a type of
vehicle. No animal
                is a car. Therefore, no
animal can be a vehicle.",
  "validity": false,
  "plausibility": true
}
```

2. After Structural Enhancement: The system extracts the logic schema into a `response` object, identifying the Subject, Predicate, and Quantifier

for each premise.


```
{
  "schema": {
    "premises": [
    {"original_index": 0, "quant": "All",
       "subj": "car", "pred": "vehicle"},
     {"original_index": 1, "quant": "No",
        "subj": "animal", "pred": "car"}
    ],
    "conclusion": {
       "quant": "No", "subj": "animal",
"pred": "vehicle"
    }
  }
}
```


3. After Noise Injection (English Noisy): Irrelevant premises (e.g., about soda and celestial bodies) are randomly injected. The `relevant_premises` array records the indices of the core logical components.


```
{
  "augmented_syllogism": "Nothing that is
a soda is a juice.
       All cars are a type of vehicle.
Anything that is a sun
      is a celestial body. No animal is
a car. Therefore,
      no animal can be a vehicle.",
  "relevant_premises": [6, 7],
  "validity": false
}
```


4. After Multilingual Expansion (Multilingual Noisy): The entire text is translated into a target language (e.g., French) while maintaining the original schema and validity labels, ensuring the model learns content-independent formal reasoning.


```
{
  "lang": "French",
  "new_syllogism": "Rien de ce qui est un
soda n'est un jus.
     Toutes les voitures sont un type de
véhicule. Tout ce
        qui est un soleil est un corps
céleste. Aucun animal
      n'est une voiture. Par conséquent,
aucun animal ne
      peut être un véhicule."
}
```