RvH-40 at SemEval-2026 Task 11: Disentangling Reasoning from Belief through Symbolic Abstraction

Janiek de Rijke
University of Groningen
j.de.rijke.1@student.rug.nl

Niek Biesterbos
University of Groningen
n.a.biesterbos@student.rug.nl

Mark den Ouden
University of Groningen
m.e.den.ouden@student.rug.nl

Abstract

Large Language Models (LLMs) often struggle with syllogistic reasoning due to "belief bias," where semantic world knowledge overrides formal logical structure. In this paper, we present our submission for the SemEval-2026 Task 11 shared task. We investigate the discrepancy between a model's latent logical capabilities and its performance on natural language text. By employing symbolic transformations, specifically variable and pseudoword substitution, we demonstrate that models like Qwen2.5-14B possess strong inherent reasoning skills that are suppressed by linguistic content. We propose a "logic alignment" strategy using Low-Rank Adaptation (LoRA) to bridge this gap. Our final model achieved a near-perfect accuracy of 97.92% on the validation set and 96.34% on the official hidden test set, effectively eliminating content bias while maintaining robust generalization across abstract formats.

1 Introduction

Syllogistic reasoning remains a critical benchmark for evaluating the formal capabilities of Artificial Intelligence, as Large Language Models (LLMs) frequently exhibit "reasoning biases" by defaulting to heuristics rather than abstract logical rules (Ozeki et al., 2024; Bertolazzi et al., 2024). This *content effect* occurs when a model's judgment of validity (the logical form of a syllogism) is distorted by real-world plausibility. In this work, we participate in subtask 1 of the SemEval-2026 Task 11 shared task (Valentino et al., 2026), centering on the hypothesis that this reasoning gap stems from a failure to decouple formal structure from semantic priors rather than a lack of innate logical capacity.

Our strategy is two-sided: first, we use symbolic and synthetic abstractions, variable and pseudoword substitution to expose the model's latent logical engine. Second, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to fine-tune the *Qwen2.5-14B-Instruct* model. This "logic alignment" encourages the model to attend to quantifiers and structural dependencies rather than lexical plausibility.

Our participation led to a significant breakthrough, with zero-shot accuracy jumping from 67.50% to 97.92% on standard syllogisms, while the Total Content Effect (TCE) was reduced from 44.54 to 1.04. Qualitatively, our system mastered complex negation structures (Group G6) and maintained high accuracy on unseen pseudoword formats, demonstrating robust, content-independent reasoning. Our code is available at: https://github.com/JDRijke/Shared_Task.

2 Background

Syllogistic inference is a robust benchmark for evaluating whether Large Language Models (LLMs) perform genuine deductive reasoning or rely on statistical heuristics. Recent research highlights a persistent "belief bias" in LLMs, where semantic plausibility often overrides formal logical validity (Ozeki et al., 2024; Bertolazzi et al., 2024).

2.1 Observed Biases and Mechanistic Insights

LLM reasoning is often divided into **content-based** (prioritizing real-world plausibility) and **form-based** reasoning (relying on abstract structure). Failures in formal reasoning are frequently attributed to the *atmosphere effect*, *conversion errors*, and *surface-level pattern matching* (Bertolazzi et al., 2024; Ozeki et al., 2024).

Mechanistic interpretability has recently shed light on these failures. Kim et al. (2025) identified specific "reasoning circuits" within LLMs responsible for syllogistic inference, suggesting that logical errors occur when semantic activations interfere with these internal structural pathways. To counteract this, Valentino et al. (2025) demonstrated that fine-grained activation steering can mitigate con-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 451–456
July 3-4, 2026 ©2026 Association for Computational Linguistics

tent effects by guiding the model's internal states toward logical form rather than belief-based priors.

2.2 Hybrid Approaches to Disentanglement

While LLMs struggle with abstract form, recent frameworks like QuaSAR (Ranaldi et al., 2025) and the use of quasi-symbolic abstractions (Xu et al., 2024) show that bridging the gap between natural language and symbolic representation can encourage reasoning aligned with logical form. Our work builds on these insights, exploring how symbolic fine-tuning can selectively inhibit belief bias and align the model's linguistic output with its latent deductive capabilities (Kim et al., 2025).

3 System Overview

Our system is designed to evaluate and enhance the logical reasoning capabilities of LLMs by isolating the abstract logical form from semantic content. We implement a multi-staged approach consisting of prompt engineering, symbolic transformation, and parameter-efficient fine-tuning.

3.1 Prompting Strategies

We evaluate several prompting configurations to obtain better reasoning:

- **Zero-shot (Normal):** Direct validity prediction without additional context.
- **Few-shot:** Providing labeled examples to ground the model in the task requirements.

3.2 Symbolic and Synthetic Abstractions

To mitigate the content bias identified in our literature review, we developed two preprocessing techniques to obscure real-world meaning:

- **Variable Substitution:** We identify category terms using Spacy's POS-tagging and replace them with abstract variables (e.g., P, Q, R). This forces the model to rely on structural dependencies rather than lexical plausibility.
- **Pseudoword Substitution:** Terms are replaced with syntactically similar but semantically null pseudowords (e.g., "mammal" becomes "alaalg"), maintaining the grammatical structure while removing common-sense cues.

3.3 Fine-Tuning with LoRA

To further specialize our models, we use Low-Rank Adaptation (LoRA) (Hu et al., 2022). The base models were fine-tuned on a subset of the training data, utilizing the different syllogistic representations (standard, variable-based, and pseudoword-based) to encourage the model to learn the underlying rules of deduction rather than surface-level pattern matching.

4 Experimental setup

4.1 Data Splits

For our experiments, we utilized the official SemEval-2026 Task 11 dataset. For the fine-tuning phase, we applied a 90/10 split on the training data, resulting in a training set of 90% to update the model weights and an internal validation set of 10% to monitor convergence and prevent overfitting. The final evaluation was performed on the official test set provided by the shared task organizers.

4.2 Preprocessing and Abstraction

As described in the System Overview, all syllogisms underwent preprocessing using the `spaCy` NLP library (version 3.4.1)[1].

- **Variable Swapping:** We implemented a custom pipeline using `en_core_web_sm` to identify nouns and noun phrases that function as category terms. These were mapped to a fixed set of variables $\{P, Q, R, X, Y, Z, U, V, W\}$ to ensure consistency across the three statements of the syllogism.
- **Pseudoword Substitution:** A character-level substitution cipher was applied to the identified lemmas, based on the letters of those lemma's.

4.3 Hyperparameters and Training

Fine-tuning was conducted using Low-Rank Adaptation (LoRA) on the *Qwen2.5-14B-Instruct* and *Llama-3.1-8B-Instruct* models. We utilized a Rank (r) of 16, alpha (α) of 32, and a dropout rate of 0.05, targeting the linear layers for adaptation. Training spanned 7 epochs with a learning rate of 2×10^{-5} and a cosine scheduler. To accommodate hardware constraints, we employed 16-bit floating-point precision (fp16) and a batch size of 1 with gradient

[1] https://spacy.io/

Format	Syllogism Example
Standard	Anything that is a bird is a vertebrate. Every bat is a vertebrate. Consequently, no bat is a bird.
Variable	Anything that is a P is a Q. Every R is a Q. Consequently, no R is a P.
Pseudoword	Anything that is a tari is a verbetrdbe. Every tdb is a verbetrdbe. Consequently, no tdb is a tari.

Table 1: Examples of the three representational formats used in our experiments. While the logical structure remains identical (Mood: AEE-2), the semantic content is progressively abstracted to isolate formal reasoning from belief bias.

accumulation steps set to 8, effectively yielding a batch size of 8.

4.4 Logical Taxonomy

To evaluate structural complexity beyond simple accuracy, we categorized the evaluation set into ten distinct taxonomic groups ($G1$–$G10$) based on the logical mood of the syllogisms. This classification is determined by two primary variables: the number of universal quantifiers (quantifying over all members of a set) and the frequency of negative operators (negating a relationship between sets).

The taxonomy is structured as follows:

Group	Universals	Negatives
G1	0	0
G2	0	1
G3	0	$\geq$2
G4	1	0
G5	1	1
G6	1	$\geq$2
G7	2	$\geq$1
G8	3	0
G9	2	0
G10	3	$\geq$1

Table 2: Logical taxonomy of the evaluation set categorized by the count of universal quantifiers and negative operators.

4.5 Evaluation Metrics

We evaluated our models using three primary metrics:

- **Accuracy (ACC):** The percentage of correctly predicted validity labels (true/false).
- **Total Content Effect (TCE):** A metric measuring the difference in accuracy between syllogisms where logic and plausibility align versus where they conflict.
- **Plausibility-Resistant Efficiency (PRE):** The primary ranking metric, calculated as $PRE = ACC/(1 + \ln(1 + TCE))$, which rewards high accuracy while penalizing sensivity to content bias.

Evaluation was performed using the `scikit-learn` library.

4.6 Zero-shot Baseline

To establish a performance baseline, we evaluated the zero-shot capabilities of two state-of-the-art open-weights models: *meta-llama/Llama-3.1-8B-Instruct* and *Qwen/Qwen2.5-14B-Instruct*. We employed a strict system prompt to enforce the required output format:

```
"You are a reasoning model. Decide whether the following syllogism is logically valid. Answer ONLY with 'true' or 'false' on a single line."
```

As shown in Table 3, both models show moderate performance on the standard dataset, with Qwen2.5-14B slightly outperforming Llama-3.1-8B (67.5% vs. 64.2%). However, both models exhibit significant biases when confronted with plausible but logically invalid statements, highlighting the necessity for our proposed fine-tuning and abstraction interventions.

Model	Accuracy
meta-llama/Llama-3.1-8B-Instruct (zero-shot)	64.17
Qwen/Qwen2.5-14B-Instruct (zero-shot)	**67.50**

Table 3: Zero-shot baseline accuracy on the standard syllogism evaluation set.

5 Results

We conducted our primary analysis on the official development test set provided during the task's training phase. To thoroughly evaluate the models' formal reasoning capabilities, we tested them across three distinct representational formats: standard natural language (*Std*), variable-based abstraction (*Var*), and character-level pseudoword substitution (*Sub*). Finally, to validate our findings, we report the official performance of our best configuration on the unseen test set via CodaBench.

5.1 Overall Performance

Table 4 presents the quantitative results across all configurations.

Baselines: The Logic Gap The *Qwen2.5-14B* base model demonstrates strong latent logical capabilities that are suppressed by linguistic content. While it achieves only 67.50% accuracy on standard text due to high content bias ($TCE = 44.54$), its performance jumps to 82.92% on abstract variables with a minimal content effect ($TCE = 3.07$). This confirms that the base model suffers from "belief bias" rather than a lack of reasoning ability. In contrast, the *Llama-3.1-8B* baseline fails to benefit from abstraction, indicating a fundamental lack of formal reasoning skills.

Impact of Fine-Tuning Our optimized fine-tuning strategy (*FT-Std*) yields a massive performance breakthrough. The model achieves near-perfect accuracy on standard syllogisms (97.92%) and effectively eradicates content bias ($TCE = 1.04$), surpassing even the abstract baselines in logical consistency.

Crucially, this robustness generalizes to unseen abstract formats. The *FT-Std* model achieves 92.92% accuracy on pseudoword substitutions (*Sub*), significantly outperforming the model trained explicitly on variables (*FT-Var*). This suggests that our training strategy has successfully aligned the model's natural language processing with its internal logical engine, rather than simply teaching it a template-matching trick.

Codabench Hidden Test Set Performance To confirm the robustness of our approach, we submitted our best-performing model, *Qwen (FT-Std)*, to the official CodaBench evaluation platform. Evaluated on this test set, our system achieved an outstanding accuracy of 96.34% and effectively suppressed semantic interference with a minimal content effect ($TCE = 3.13$). This resulted in a final combined primary ranking score (PRE) of 39.86, placing our team, *RvH40*, at 25^{th} position on the global leaderboard.

While the top of the leaderboard reached a perfect score of 100.0, it is important to note that our result was achieved using a single, parameter-efficiently fine-tuned 14B model without the use of complex ensembles or larger-scale architectures. These official metrics closely align with the development validation results, demonstrating that our strategy successfully generalized the underlying rules of syllogistic logic without overfitting to the development data.

Model	Format	ACC	TCE	PRE
Baselines				
Llama-3.1-8B	Std	64.17	32.15	14.26
	Var	60.83	14.82	16.17
	Sub	58.75	8.45	18.10
Qwen2.5-14B	Std	67.50	44.54	14.01
	Var	82.92	3.07	34.49
	Sub	82.08	4.76	29.85
Fine-tuned				
Qwen (FT-Std)	Std	**97.92**	**1.04**	**57.09**
	Var	92.50	6.25	31.03
	Sub	92.92	4.03	35.54
Qwen (FT-Var)	Std	69.17	37.96	14.83
	Var	83.33	4.49	30.82
	Sub	84.58	7.64	26.79

Table 4: Results on the development validation set. *Qwen FT-Std* demonstrates superior performance across all metrics, achieving the highest accuracy and lowest content bias even compared to abstract baselines.

6 Error Analysis

To investigate the mechanisms driving the performance gains, we conducted an error analysis focusing on two dimensions: (1) the interaction between logical validity and belief bias, and (2) the impact of structural complexity (negations and quantifiers).

6.1 Decoupling Logic from Belief

A robust reasoning model should exhibit performance invariance across semantic conditions. Figure 1 visualizes this interaction effect, comparing the standard models against their abstract counterparts. A `Conflict` condition occurs when logical validity contradicts real-world plausibility, whereas a `Consistent` condition is one where logic and belief align.

The *Qwen (Base)* model on standard text (dark red dashed line) displays a classic "belief bias" signature: high accuracy on consistent items but a sharp drop in conflict scenarios ($< 50\%$), where it defaults to world knowledge. However, when the same model is tested on *Variable* or *Pseudoword* data (lighter red/pink lines), the performance lines flatten and jump to $> 80\%$ accuracy. This confirms that the reasoning engine is intact but suppressed by semantic interference.

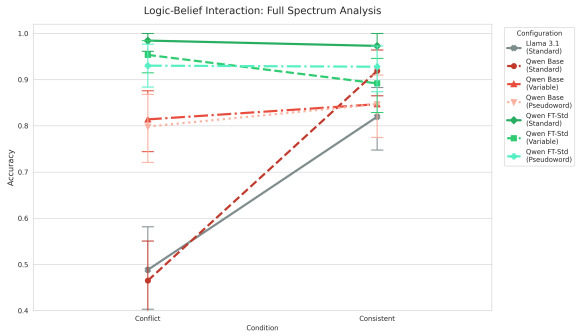

Figure 1: Logic-Belief Interaction Plot. The baseline models (Llama and Qwen Base in red/grey) show a steep slope, indicating failure in conflict scenarios. The abstract baselines (Qwen Base Var/Sub) show a flat, high trajectory, revealing latent logical capability. Our fine-tuned model (Qwen FT-Std in green) successfully replicates this flat trajectory across all formats, proving that reasoning has been decoupled from semantic belief.

Crucially, our *FT-Std* model (solid green line) aligns perfectly with these abstract baselines, maintaining near-perfect accuracy ($> 97\%$) regardless of plausibility. This demonstrates that fine-tuning did not merely teach the model to ignore content, but effectively mapped the robust logical operators available in the abstract latent space to the natural language domain.

6.2 Structural Competence and Negation

We further decomposed performance into ten structural groups ($G1$–$G10$) to identify specific logical blind spots. Figure 2 presents a comparative heatmap of accuracy across model configurations.

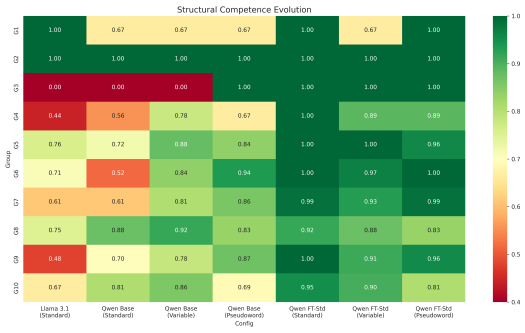

Figure 2: Structural Competence Heatmap. Note the progression of Group G6 (multiple negations) from red in the base model (0.52) to dark green in the fine-tuned model (1.00).

The Negation Breakthrough (G6) The most significant finding concerns Group 6 ($G6$), characterized by syllogisms with a single universal premise and multiple negations (e.g., *"No A are B, Some C are A"*). While the base model fails catastrophically on standard text ($Accuracy = 0.52$)—likely due to the cognitive load of tracking overlapping negative scopes—its latent capability is proven by excellent performance on pseudowords ($Accuracy = 0.94$). Our fine-tuned model (*FT-Std*) successfully bridges this gap, achieving perfect accuracy (1.00) on G6 in standard text. This transfer success indicates that the training signal allowed the model to leverage its internal "pseudoword logic" for processing natural language negations.

Our findings indicate that LLMs possess latent structural reasoning beyond statistical pattern matching. The high zero-shot performance of Qwen Base on variables (82.9%) compared to standard text (67.5%) suggests that logical operators are present but suppressed by semantic priors. Rather than teaching new data, fine-tuning "aligned" linguistic processing with this logical core. The model's generalization to unseen pseudowords (92.9%) confirms it attends to argument structure—quantifiers and negations—rather than content. This supports using LoRA to selectively inhibit belief bias while maintaining robust reasoning capabilities.

7 Conclusion

In this study, we demonstrated that the logical reasoning gap in LLMs is largely a product of content bias rather than a lack of innate capability. We showed that the Qwen2.5-14B model holds strong latent logic, which we successfully activated for natural language using parameter-efficient fine-tuning (LoRA). Our final model achieved 97.9% accuracy on the validation set and 96.3% on the official hidden test set, effectively solving the "belief bias" problem for syllogistic reasoning. The model's ability to generalize this logic to abstract and pseudoword formats confirms that it has learned a robust, content-independent reasoning strategy. Future research could explore whether this "logic alignment" approach generalizes to more complex logical forms beyond the syllogistic domain.

8 Limitations

While our system achieves near-perfect accuracy on the provided syllogistic dataset, several limitations remain. First, our fine-tuning approach was specifically optimized for categorical syllogisms

consisting of two premises and a single conclusion. It remains to be seen whether the "logic alignment" observed here generalizes to more complex logical structures, such as polysyllogisms or non-categorical deductive reasoning.

Second, our symbolic transformations (variable swapping and pseudoword substitution) rely on the accuracy of the underlying POS-tagging and lemmatization provided by the `spaCy` library. Errors in term identification could lead to inconsistent abstractions, potentially impacting the model's performance on noisier, real-world data.

Finally, our study focused on English-language syllogisms. While the principles of formal logic are universal, the linguistic cues and quantifiers in other languages may present different challenges for belief bias mitigation that were not explored in this work.

9 Acknowledgments

We would like to express our gratitude to Prof. M. (Malvina) Nissim and H. (Huiyuan) Lai, MA, both from the University of Groningen, for their insightful guidance and support throughout this project. We are grateful for their expertise in natural language processing and logical reasoning, which helped us significantly in refining our experiments and analysis. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster. Additionally, we thank the SemEval-2026 Task 11 organizers for providing the datasets and the CodaBench platform for official evaluation, which enabled us to validate our system's performance on a global scale.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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.

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, Bangkok, Thailand. Association for Computational Linguistics.

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. *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)*.

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.