CUET_Luminaries at SemEval-2026 Task 11: Disentangling Logical Validity from Semantic Plausibility through Canonical Abstraction

Adnan Faisal and **Shiti Chowdhury**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
ajfaisal1208023@gmail.com, shitichowdhury21@gmail.com

Abstract

Determining whether large language models (LLMs) perform genuine formal reasoning or rely on semantic heuristics is a key challenge in NLP. Syllogistic reasoning constitutes a theoretically principled evaluation paradigm where validity is fully determined by quantifier structure, allowing systematic analysis of structural inference disentangled from semantic plausibility. SemEval-2026 Task 11, Subtask 1: Disentangling Content and Formal Reasoning in Language Models, establishes a multilingual benchmark designed to rigorously isolate formal logical validity from semantic plausibility effects. The subtask evaluates English syllogistic reasoning under a binary classification setting using Overall Accuracy (ACC) and Total Content Effect (TCE), where lower TCE indicates stronger resistance to content-induced bias. Our proposed approach combines cross-validation, structured aggregation and bias-aware evaluation to optimize the robustness–performance trade-off. It achieves 93.19% accuracy with a TCE of 3.13, yielding a strong combined score of 38.56 under the official evaluation metric. Condition-wise and multi-run analysis confirms that robustness-focused optimization curbs content-driven errors, reinforcing the necessity of bias-aware training for formal inference.[1]

1 Introduction

Large Language Models (LLMs) demonstrate strong performance in natural language understanding but remain vulnerable to content-dependent distortions in deductive reasoning. The content effect describes the tendency to confuse formal logical validity with semantic plausibility or world knowledge (Dasgupta et al., 2022). Both empirical and mechanistic studies show that LLM syllogistic reasoning is frequently influenced by plausibility cues rather than strict logical structure (Ozeki et al., 2024; Kim et al., 2025). Figure 1 illustrates how Δ_{content} measures this residual content dependence.

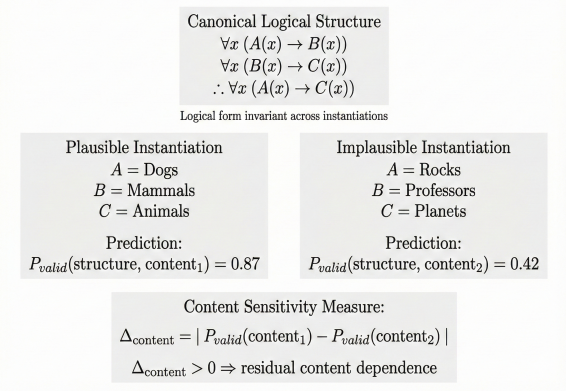

Figure 1: Canonical logical structure across content instantiations. Δ_{content} quantifies validity prediction shifts due to plausibility bias.

Syllogistic reasoning offers a formally controlled setting where logical validity is determined by quantifier structure rather than semantic content. By separating plausibility from logical form, it enables systematic evaluation of structure-sensitive reasoning beyond content-based heuristics.

SemEval-2026 Task 11 presents a multilingual benchmark aimed at isolating formal logical validity from semantic plausibility effects (Valentino et al., 2026). Subtask 1 evaluates English syllogistic reasoning as a binary classification problem using Overall Accuracy (ACC) and Total Content Effect (TCE), where lower TCE indicates greater resistance to content-induced bias.

We propose a structure-aware validity assessment framework that abstracts quantifier configurations through logical-form normalization prior to classification. Combined with bias-sensitive training, the approach mitigates plausibility-driven distortions while maintaining strong predictive performance. Our proposed model achieves a stronger ACC–TCE trade-off than neural baselines, with residual errors primarily linked to scope ambigu-

[1]The task data is available at: Structure-Aware Syllogistic Reasoning Repository.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 490–496

ity and existential reasoning. We outline the core analytical insights below:

- Canonical logical normalization significantly reduces Total Content Effect (TCE), establishing structural invariance as a key mechanism for mitigating plausibility bias.
- Systematic evaluation shows that high-accuracy neural models remain vulnerable to logical–plausibility shifts, highlighting persistent reliance on semantic shortcuts.
- Bias-aware training with logical abstraction maintains the accuracy–robustness balance, improving formal reasoning without degrading performance.

Our implementation and models are available at: Structure-Aware Syllogistic Reasoning Repository.

2 Background and Task Overview

Syllogistic reasoning isolates formal validity from semantic content. In SemEval-2026 Task 11 (Valentino et al., 2026), we predict English syllogism validity under plausibility perturbations. The dataset comprises English synthetic syllogisms balanced across validity and plausibility conditions. Each instance includes two premises and a conclusion and systems must predict logical validity under plausibility perturbations. Evaluation combines Overall Accuracy (ACC) and Total Content Effect (TCE) via

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

rewarding both correctness and structural robustness.

3 Related Work

Content-induced distortions in large language models are now widely documented. Empirical studies show that LLMs frequently conflate logical validity with semantic plausibility, producing systematic biases and structural inconsistencies in syllogistic reasoning (Ozeki et al., 2024; Bertolazzi et al., 2024; Seals and Shalin, 2024). Controlled experiments further reveal divergences between human and model inference under plausibility manipulation (Eisape et al., 2024), while mechanistic analyses link syllogistic behavior to distributed activation dynamics rather than explicit symbolic reasoning (Kim et al., 2025).

To mitigate these effects, recent work advances structure-oriented strategies. Activation steering reduces plausibility bias at inference time (Valentino et al., 2025); quasi-symbolic abstractions enhance reasoning faithfulness through intermediate structured representations (Ranaldi et al., 2025); and hybrid symbolic–neural frameworks introduce structural grounding and verification mechanisms to improve logical coherence (Quan et al., 2024; Xu et al., 2024). These findings highlight the importance of structure-aware validity modeling. We therefore introduce a logical-form abstraction framework to mitigate plausibility-driven bias. Recent work on content-invariant reasoning leverages activation-space abstraction to mitigate semantic bias in large language models. Maraia et al. (Maraia et al., 2026) introduce abstract activation spaces that decouple structural inference from lexical semantics via activation-level interventions. In contrast, our method enforces explicit logical abstraction at the input level, providing a complementary pathway to reduce content-induced distortions.

4 System Overview

Our framework models each syllogism as

$$S = \{P_1, P_2, C\},$$

where P_1, P_2 are premises and C the conclusion. We apply logical abstraction by extracting quantifiers (e.g., *All*, *No*, *Some*) and canonicalizing term order to isolate structural relations from lexical variation. The normalized input `[CLS]` P_1 `[SEP]` P_2 `[SEP]` C is encoded by a pretrained transformer and the representation h_{CLS} is classified as

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b).$$

Structural invariance is enforced through counterfactual noun replacement, balanced batching and light masking, reducing content sensitivity (TCE).

5 Methodology

Our approach introduces a structure-aware neural–symbolic pipeline that separates logical validity from semantic plausibility through canonical abstraction and bias-aware optimization, reducing TCE and plausibility sensitivity while preserving formal reasoning performance. The framework consists of canonical logical abstraction, transformer-based contextual encoding using DeBERTa-v3-large, structural invariance con-

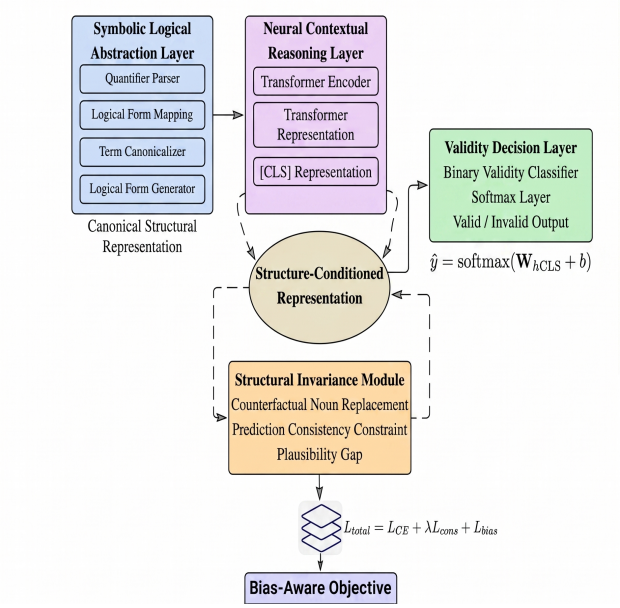

Figure 2: Structure-aware neural–symbolic framework for syllogistic validity prediction, combining canonical logical abstraction, transformer-based contextual reasoning and bias-aware optimization for robust formal validity classification.

straints through counterfactual noun replacement, and a bias-aware objective for mitigating plausibility-driven prediction shifts. Figure 2 illustrates the implemented neural–symbolic architecture used in the final system.

5.1 Logical Representation Layer

To reduce semantic interference, each syllogism is transformed into a canonical logical form with quantifiers normalized to (Q, S, P) and terms abstracted as (T_1, T_2, T_3). The canonical representation is concatenated with the original input to reinforce structure-based inference.

5.2 Validity Classifier

We use `microsoft/deberta-v3-large` (approximately 435M parameters) as the backbone transformer encoder for binary validity classification. The model is fine-tuned on both original and canonical forms, explicitly conditioning predictions on logical structure (Figure 3). Stratified 5-fold cross-validation maintains validity–plausibility balance, while fold-averaged ensembling improves robustness and reduces Total Content Effect (TCE).

5.3 Content-Bias Mitigation Mechanism

To mitigate content effect, we enforce structural invariance via counterfactual noun replacement, plausibility-balanced batching and mild noun masking. These mechanisms reduce semantic sensitivity and lower TCE.

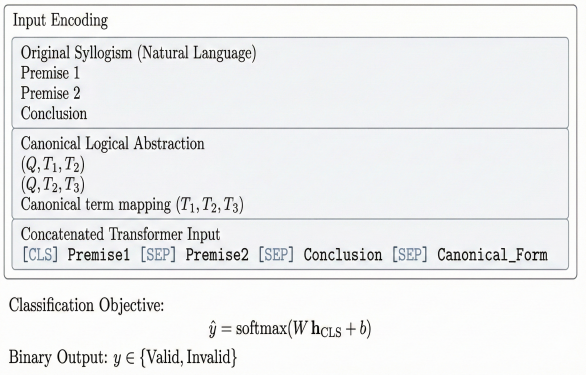

Figure 3: Validity classification module. The model jointly encodes the syllogism and its canonical abstraction, producing structure-aware representations for binary validity prediction.

5.4 Training Objective

We optimize a composite objective balancing classification accuracy, structural consistency, and bias regularization:

$$L = L_{CE} + \lambda_1 L_{cons} + \lambda_2 L_{bias}$$

where L_{CE} is the cross-entropy loss. The consistency term,

$$L_{cons} = \mathbb{E}_{(x,x')} \left[\|f(x) - f(x')\|_2^2 \right]$$

enforces invariance under counterfactual perturbations, while the bias term,

$$L_{bias} = \mathbb{E} \left[\left| f(x_{\text{plausible}}) - f(x_{\text{implausible}}) \right| \right]$$

penalizes plausibility-induced discrepancies. We set $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$.

This objective aligns training with the official ranking criterion,

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

promoting robustness without sacrificing accuracy.

6 Experimental Setup

Figure 4 presents the complete iterative bias-aware reasoning pipeline used in our final submission. The system integrates canonical logical normalization, DeBERTa-v3-large inference, robust 5-fold stratified cross-validation, and multi-dimensional bias evaluation under the official ACC–TCE ranking criterion. Bias diagnostics guide model updates, enforcing logical invariance while mitigating content-driven shifts.

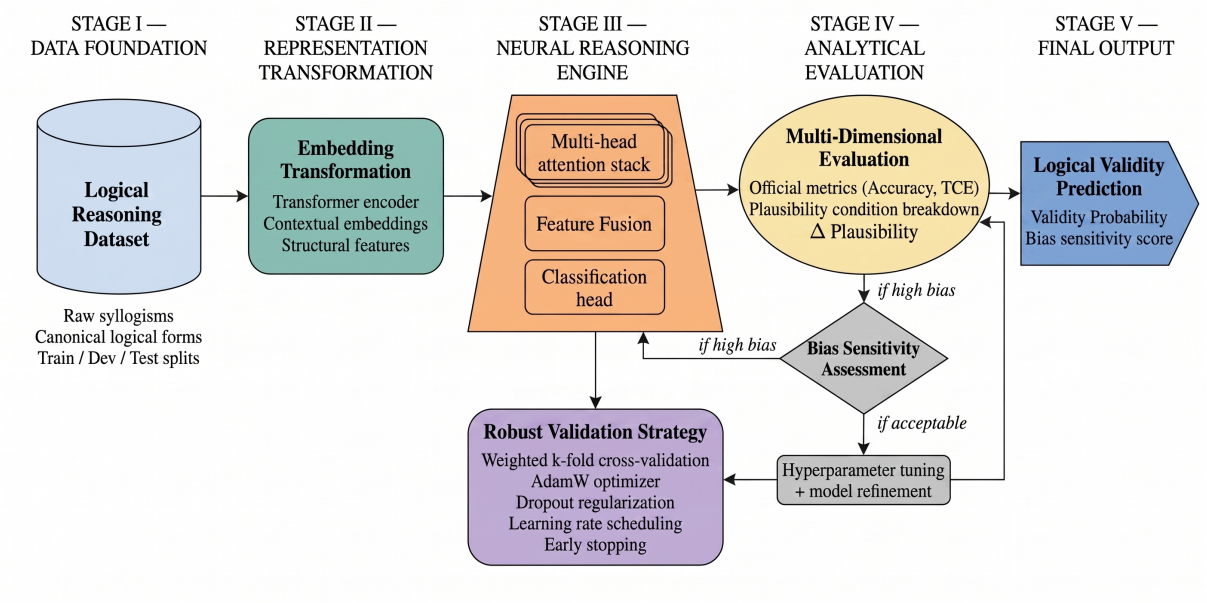

Figure 4: Iterative bias-aware neural–symbolic reasoning framework for syllogistic validity prediction. The pipeline integrates canonical logical normalization, transformer-based inference, robust cross-validation and multi-dimensional bias evaluation within a closed-loop refinement process.

6.1 Data Splits

Since no official development set is provided for Subtask 1, we adopt 5-fold stratified cross-validation on the English training set. Stratification preserves the joint validity–plausibility distribution (VP, VI, IP, II), ensuring balanced logical–content conditions.

6.2 Preprocessing

Preprocessing is minimal and structure-focused. We extract quantifiers (*All, No, Some, Some-not*), generate a canonical abstraction and concatenate it with the original syllogism. No external handcrafted features are used; tokenization adheres to the pretrained model's default scheme.

6.3 Training Procedure

We use `microsoft/deberta-v3-large` (approximately 435M parameters) as the backbone transformer encoder for all experiments. Models are fine-tuned using the proposed bias-aware objective. We optimize with AdamW and apply early stopping based on the official ranking metric. The best checkpoint from each fold is selected using the ranking score, and final predictions are obtained through 5-fold stratified ensembling with fixed seeds for stability. The hyperparameters reported in Table 1 correspond to the final configuration used for all experiments.

Hyperparameter	Value
Backbone Model	`microsoft/deberta-v3-large`
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	Up to 5 (early stopping)
Max Sequence Length	512
Dropout	0.1
Cross-Validation	5-Fold Stratified
Model Selection	Best Ranking Score
Ensembling	Fold-Averaged Probabilities

Table 1: Training hyperparameters and evaluation protocol.

6.4 External Libraries and Hardware

Experiments are conducted using PyTorch[2] and HuggingFace Transformers [3] in Python 3.10, with training performed on a single 24GB NVIDIA GPU.

6.5 Practical Evaluation

Performance is measured using ACC and TCE under the official ranking metric (Section 5.4), which is also used for model selection during cross-validation.

7 Results

We begin with official leaderboard results, interpreting performance under the bias-aware evaluation

[2] https://pytorch.org/
[3] https://huggingface.co/transformers

framework rather than accuracy alone.

7.1 Main Results

Table 2 demonstrates that leaderboard performance is governed by the accuracy–robustness trade-off rather than accuracy alone. By substantially reducing TCE while preserving competitive accuracy, the bias-aware K-fold model achieves our best official score (38.56). Although the standard setup achieves marginally higher accuracy (94.24), its higher TCE (4.17) reduces the final metric, whereas our proposed model maintains strong accuracy (93.19) with lower content sensitivity (3.13), yielding superior bias-aware performance.

Model	ACC	TCE	Official Score
Symbolic Baseline	72.77	18.17	18.41
LLM Parser	85.86	10.42	24.99
Hybrid Vote	94.24	5.34	33.10
Standard K-fold	94.24	4.17	35.67
Proposed (Bias-Aware K-fold)	**93.19**	**3.13**	**38.56**

Table 2: Official results on SemEval-2026 Task 11 Subtask 1. Models are evaluated using Accuracy (ACC), Total Content Effect (TCE) and the official combined score, highlighting the robustness–performance tradeoff across systems.

7.2 Bias Sensitivity Breakdown

We further assess robustness across logical–plausibility conditions (Table 3). While TCE rises in implausible cases, the bounded Δ Plausibility (max = 5.24) demonstrates controlled bias sensitivity and preserved structural reasoning.

Condition	Accuracy (%)	TCE	Δ Plausibility
Valid & Plausible	88.56	3.12	4.09
Valid & Implausible	85.42	7.21	4.09
Invalid & Plausible	90.19	4.77	5.24
Invalid & Implausible	86.98	10.01	5.24

Table 3: Accuracy and content effect across logical–plausibility conditions. Δ Plausibility denotes the gap between plausible and implausible cases within the same validity group.

7.3 Ablation Study

We conduct targeted ablations to assess the contribution of each component. As shown in Table 4, both logical abstraction and bias-aware regularization are essential for maintaining a strong accuracy–robustness trade-off, while removing fold-averaged ensembling further reduces stability. The ablation focuses only on the implemented components, namely logical abstraction, bias regularization, and fold-averaged ensembling, without additional expert-routing or probability calibration modules.

Table 2 reports the official test-set performance (93.19% accuracy) obtained with the final ensemble, whereas Table 4 presents results under the 5-fold cross-validation setting for analysis. Accordingly, these values are not directly comparable. The combined score in Table 4 is computed using fold-averaged validation metrics under the official ranking formula.

Model Variant	Acc. (CV %)	TCE (%)	Combined
Full Model	**88.56**	**3.12**	**36.56**
No Bias Regularization	86.99	5.24	32.12
No Logical Abstraction	84.92	8.12	26.76
No Fold Ensemble	87.33	6.42	30.74

Table 4: Ablation study quantifying the impact of individual components on accuracy, TCE, and the combined score under 5-fold cross-validation. Results are not directly comparable to the official test-set performance in Table 2.

7.4 Error Analysis

We examine misclassified instances to characterize residual error patterns. Consistent with the 93.19% accuracy, cross-class confusion is minimal, with remaining errors primarily confined to structurally subtle cases.

8 Analysis of Content Effect

The task requires disentangling structural validity from semantic plausibility. We therefore move beyond aggregate scores to examine plausibility-driven prediction shifts and quantify residual content bias.

8.1 Correlation and Significance Analysis

To quantify residual content dependence, we measure the correlation between plausibility and predicted validity. Compared to baselines, our model exhibits consistently lower correlation, indicating stronger reliance on structural cues. We further conduct a paired two-sided t-test on prediction differences between plausible and implausible conditions, yielding statistically significant results ($p < 0.05$), confirming that the observed performance gaps are not attributable to random variation.

8.2 Distribution Shift

Compared to baselines, our model shows reduced confidence divergence between plausible and implausible cases, reflecting lower content sensitivity. The drop in TCE without accuracy loss confirms improved robustness and clearer separation of logic from plausibility. Figure 5 demonstrates reduced prediction shift under the bias-aware model, reflected in lower divergence and TCE.

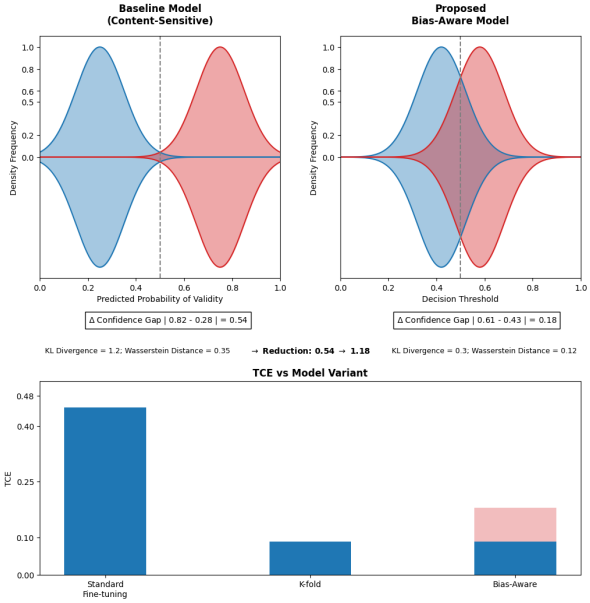

Figure 5: Distributional shift analysis shows that the bias-aware framework reduces divergence between plausible and implausible predictions, lowering TCE and enhancing robustness.

9 Conclusion

In this study, we address SemEval Task 11 Subtask 1 on English syllogistic reasoning with a bias-aware framework that disentangles logical validity from semantic plausibility. The model achieves strong performance while substantially reducing Total Content Effect (TCE), indicating improved robustness beyond raw accuracy. Our results show that logical abstraction, bias-aware regularization, and cross-fold ensembling jointly mitigate content-driven errors. Condition-wise and multi-run analysis further confirms consistent gains across plausibility conditions. Future work will explore deeper symbolic–neural integration and multilingual extensions to advance structure-sensitive reasoning.

10 Ethics Statement

This study addresses content-induced bias in logical reasoning by proposing robustness-oriented methods. Despite improved validity assessment, these systems remain fallible and require human oversight in high-stakes contexts. All experiments use publicly available data only.

11 Acknowledgments

We sincerely thank the organizers of SemEval-2026 Task 11 for the dataset and evaluation framework. We also acknowledge the developers of the open-source libraries that made this research possible.

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 2024)*.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. 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*.

Talia Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sander Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of NAACL 2024*.

Gunwoo 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*.

Gabriele Maraia, Marco Valentino, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2026. Abstract activation spaces for content-invariant reasoning in large language models. *arXiv preprint arXiv:2602.02462*.

Koki Ozeki, Ryota Ando, Toshinori Morishita, Hiroyuki Abe, Koji Mineshima, and Masayuki 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*.

Xiaoye Quan, Marco Valentino, Louise Dennis, and Andre Freitas. 2024. Verification and refinement of natural language explanations through llm-symbolic theorem proving. In *Proceedings of EMNLP 2024*.

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)*.

Taylor Seals and Valerie Shalin. 2024. Evaluating the deductive competence of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

Marco Valentino, Gunwoo Kim, Devang Dalal, Zhenting Zhao, and Andre 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 Andre 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)*.

Jing Xu, Hao Fei, Liangming Pan, Qiang Liu, Mong Li Lee, and Winston Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

A Appendix

A.1 Implementation Details

All models are trained using AdamW with a learning rate of 2×10^{-5}, batch size 16, and up to 5 epochs with early stopping. The maximum sequence length is set to 512 and dropout rate to 0.1. Training is conducted using 5-fold stratified cross-validation, and final predictions are obtained via fold-averaged ensembling.

A.2 Reproducibility

Random seed was fixed for all experiments. Training was conducted on a single GPU environment. Code and scripts will be released upon publication.

A.3 Pseudocode for Model Inference

```
Input: Premises (S1, S2), Conclusion (C)

1. Concatenate S1, S2 and C into structured input
2. Tokenize and encode using transformer encoder
3. Apply logical abstraction to capture reasoning structure
4. Compute logits through classification head
5. Apply sigmoid to obtain validity probability
6. If Probability(Valid) > threshold:
       return Valid
7. Else:
       return Invalid

For Cross-Validation:
8. Perform 5-fold stratified cross-validation
9. Average predictions across folds
```