# GheGheGhe at SemEval-2026 Task 11: Decoupling Logic from Belief with Bias-Targeted Fine-Tuning and Neuro-Symbolic Syllogistic Reasoning

**Gogu Răzvan-Costinel** **Plăcintescu Ștefan** **Vultur Sofia-Maria**
University of Bucharest
Faculty of Mathematics and Computer Science
{razvan.gogu, stefan.placintescu, sofia.vultur}@s.unibuc.ro

## Abstract

This paper presents a multi-paradigm approach to the first two subtasks of SemEval-2026 Task 11. For the first subtask, we explore two complementary strategies: a Llama-3 8B PEFT Majority Vote Ensemble, trained with bias-targeted augmented data, and a hybrid approach that separates LLM processing from logical reasoning, converting sentences into canonical logical forms for deterministic analysis. The hybrid approach is further extended to the second subtask. Official results placed us 17th in the first subtask and 15th in the second. Post-evaluation analysis indicates that our best model achieved perfect accuracy on the first subtask and revealed several errors in the ground truth data. After identifying certain implementation issues in the second subtask approach, the F1 retrieval score increased to over 98%, which would place us within the top 5 on the leaderboard.

## 1 Introduction

As large language models advance, they exhibit increasingly sophisticated behavior resembling structured reasoning. This progress raises a fundamental question: what are the inherent limitations of their capabilities and to what extent can they be overcome? The SemEval-2026 Shared Task 11, *Disentangling Content and Formal Reasoning in Language Models* by Valentino et al. (2026), addresses this challenge by isolating formal syllogistic reasoning from interfering content-driven biases.

Initial sensitivity analysis on zero-shot and Chain-of-Thought prompting revealed significant instability: CoT often initiated formal symbolic reasoning before undergoing a mid-generation heuristic shift, hallucinating steps toward semantically plausible conclusions. We address this limitation through two complementary approaches, detailed in our publicly available implementation.[1]

[1]https://github.com/gogurazvan/GheGheGhe-at-SemEval-2026-Task-11

In our first strategy, we investigated whether neural architectures can be calibrated toward belief-agnostic reasoning. To this end, we trained a LLaMA-3 8B parameter-efficient ensemble using a contrastively augmented curriculum to sharpen structural discrimination. We analyze representational dynamics using hidden-state subspace decomposition with PCA across three stages: (1) baseline LoRA (Hu et al., 2022) adaptation, (2) augmentation-enhanced LoRA, and (3) a LoRA majority-vote ensemble. These analyses reveal persistent overlap between semantic and logical signals in representation space. We mitigate this overlap at the decision level through logit margin calibration. By enforcing a conservative decision threshold, the model prioritizes structural validity over semantic plausibility.

Our second strategy moves away from end-to-end reasoning. Instead, we use the LLM as a linguistic encoder to map natural language into structured representations, delegating the logical inference to an explicit external reasoning component. This decoupling of language understanding from formal logic aims to reduce content-induced biases and enforce logically sound inference.

## 2 Related Work

Recent work from Dasgupta et al. (2022) shows that LLMs exhibit systematic content effects in reasoning tasks, often displaying human-like preference for semantically plausible conclusions over formally valid ones. Similarly, Eisape et al. (2024) and Ozeki et al. (2024) report consistent failures in syllogistic reasoning when plausibility conflicts with validity. Bertolazzi et al. (2024) characterize LLMs as "soft reasoners" sensitive to lexical cues even in controlled task. Seals and Shalin (2024) and Wysocka et al. (2025) further evaluate LLM deductive competence across general and biomedical syllogistic settings, consistently finding systematic



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 3112–3123

reasoning failures: limitations that motivate the SemEval-2026 Task 11 focus on quantifying intra- and cross-plausibility content effects.

Several methods have been proposed to mitigate reasoning biases in LLMs. Valentino et al. (2025) explore activation steering to reduce content effects during inference, while Ranaldi et al. (2025) introduce quasi-symbolic abstractions to improve chain-of-thought reasoning. Faithful reasoning approaches (Lyu et al. (2023); Xu et al. (2024)) incorporate symbolic constraints into generation pipelines to improve logical consistency. In contrast, our work investigates parameter-efficient fine-tuning combined with logit-level inference control and compares this neural approach with a deterministic canonical reformulation pipeline.

Understanding whether logical validity is internally encoded in language model representations is an active research area. Kim et al. (2025) analyze reasoning circuits in transformer architectures, while Maraia et al. (2026) propose abstract activation spaces for content-invariant reasoning. Our probing and logit-margin analyses extend this line of work by empirically evaluating the separability of validity and plausibility signals in fine-tuned models.

# 3 Methodology

## 3.1 Task and Data Overview

Our participation focused on the English tracks of the shared task. The training dataset consists of 960 syllogisms, each containing two premises and a conclusion. Each sample is annotated for both logical validity and semantic plausibility, enabling segmentation into four analytical quadrants: Valid–Plausible, Valid–Implausible, Invalid–Plausible, and Invalid–Implausible. Because our objective was to disentangle formal reasoning from content effects, we placed particular emphasis on the conflict cases (Valid–Implausible and Invalid–Plausible). Given our use of two distinct system paradigms, a generative LLM and a deterministic rule-based solver, we applied separate preprocessing pipelines tailored to each architecture. Both systems are evaluated using Accuracy and Content Effect (CE), the average of Intra-Plausibility CE (bias toward a validity label) and Cross-Plausibility CE (bias toward plausibility), with lower CE indicating greater reliance on logical structure.

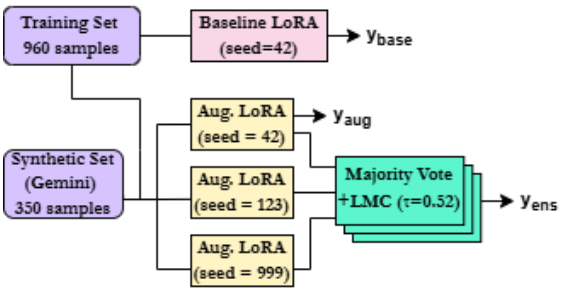


Figure 1: Neural validity pipeline. Arrows represent the inference paths for the baseline ($y_{base}$), augmented ($y_{aug}$), and final ensemble ($y_{ens}$) models.

## 3.2 Neural Validity Modeling

### 3.2.1 Chat Formatting and Label-Masked Fine-Tuning

For our generative experiments, the dataset was reformatted into a conversational chat-style structure, mapping each input syllogism to a user prompt and the corresponding validity label to the assistant's response. To optimize fine-tuning under a Causal Language Modeling objective, we applied a strict token masking strategy. All tokens corresponding to the user prompt were replaced with -100, the default `ignore_index` in PyTorch's Cross-Entropy loss implementation. This masking ensures that the loss is computed exclusively over the final classification tokens (VALID or INVALID), preventing gradient updates from being influenced by reconstruction of the input text. This prevented the model from wasting representational capacity on input reconstruction and focused the LoRA adapters purely on the logical reasoning task.

### 3.2.2 Targeted Data Augmentation

Following initial LoRA fine-tuning on the original training set, we conducted a diagnostic evaluation on a held-out validation split derived from the training data. Analysis of high-confidence misclassifications revealed persistent belief-bias failures, particularly in Valid–Implausible and Invalid–Plausible configurations. These observations motivated a targeted augmentation phase designed to strengthen structural discrimination in conflict cases, where logical mood constraints and vocabulary variation rules were enforced to ensure structural diversity and avoid repetitive patterns. To this end, we generated 350 additional syllogisms using Gemini 3.0 Flash as a synthetic data engine. The prompting strategy constrained the model to cycle through predefined classical syllogistic moods (e.g., EAE,

AII, EIO, OAO), ensuring systematic coverage of logical forms.

This process produced three categories of structurally controlled examples:

**(1) Absurd but Valid** (100 samples) – logically valid syllogisms with surreal content *(e.g., "All polka-dotted whispers are floating turnip-clouds. No sonic cacti are floating turnip-clouds. Therefore, some polka-dotted whispers are not sonic cacti.")*;

**(2) Plausible but Invalid** (110 samples) – believable conclusions that do not logically follow from the premises *(e.g., "No mammals are birds. All whales are mammals. Therefore, all whales are birds.")*;

**(3) Contrastive pairs** (140 samples) – minimally perturbed variants where a single quantifier flips validity *(e.g., **Valid:** "All $M$ are $P$. All $S$ are $M$. $\therefore$ All $S$ are $P$." vs. **Invalid:** "All $M$ are $P$. Some $S$ are $M$. $\therefore$ All $S$ are $P$.")*.

#### 3.2.3 Neural Validity Experimental Setup

**Stage 1: Baseline LoRA Adaptation** In the primary stage, we adapted Llama-3 8B using a Causal Language Modeling objective with a 512-token context window (AI@Meta, 2024). Following standard protocols (Hu et al., 2022), we applied LoRA to the attention projection layers (q_proj and v_proj) with rank $r = 16$ and scaling factor $\alpha = 32$. This configuration was selected to effectively capture the structural regularities of syllogistic reasoning while preserving the base model's pretrained linguistic capabilities.

**Stage 2: Augmentation-Enhanced LoRA** Building upon the baseline, this configuration incorporated the **contrastive augmented curriculum** described in Section 3.2.2, designed to reduce entanglement between semantic plausibility and formal validity. This targeted intervention resulted in a measurable reduction in the *Total Content Effect*.

**Stage 3: Majority Vote Ensemble** The final configuration employed a **majority vote ensemble** consisting of three independently fine-tuned LoRA adapters, each initialized with a different random seed. This ensemble strategy was adopted to stabilize high-variance decision boundaries. For a given syllogism $S$, each model $m \in \{1, 2, 3\}$ produces a binary prediction $y_m \in \{0, 1\}$. The final ensemble prediction $Y$ is defined as:

$$Y = \mathbb{I}\left(\sum_{m=1}^{3} y_m \geq 2\right), \quad (1)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

#### 3.2.4 Logit Margin Calibration (LMC)

To further refine ensemble predictions under belief-bias pressure, we implemented **Logit Margin Calibration (LMC)**. For each input, we extracted the logits corresponding to the `VALID` ($L_v$) and `INVALID` ($L_i$) tokens. Rather than applying a standard $\arg\max$, we introduced a conservative decision threshold $\tau$ on the logit margin, such that a valid verdict need to meet the condition $(L_v - L_i) > \tau$. The optimal threshold $\tau = 0.520$ was determined via grid search on the validation set, maximizing the official Combined Score by requiring the model to overcome semantic priors before endorsing a conclusion as logically valid.

### 3.3 Neuro-Symbolic Logical Solver

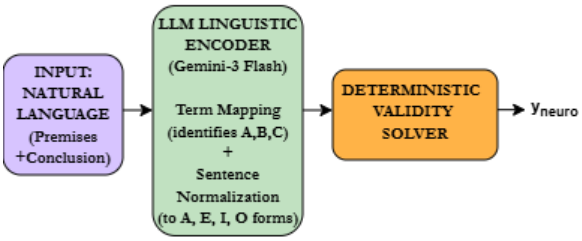


Figure 2: Neuro-symbolic pipeline for decoupling linguistic processing from formal reasoning.

#### 3.3.1 Syllogism Simplification

In the second method, we experimented with a hybrid solution. We employed LLMs to transform the syllogisms into a standard logic formula, which could later be deterministically evaluated, thus bypassing the models' plausibility biases.

To ensure bias is minimized, this simplification was performed in two steps: **term mapping**, where the model identifies and maps **major, minor, and middle terms** to identifiers $A, B$ and $C$; and **sentence normalization**, where statements are converted into one of the four standard categorical forms: **A** (**Universal Affirmative**: All $X$ are $Y$), **E** (**Universal Negative**: No $X$ are $Y$), **I** (**Particular Affirmative**: Some $X$ are $Y$), or **O** (**Particular Negative**: Some $X$ are not $Y$).

In this case, we transform a complex problem of analyzing contextual relations between sentences into a much simpler one which focuses on transforming a single sentence in one of four forms.
After the syllogisms were normalized, their validity could be deterministically evaluated using two concepts: **mood** (sequence of three letters representing the standard forms) and **figure** (defined by the position of the *middle term*). The **figure** has four forms: **Sub-Pre:** (Subject of *major*/Predicate of *minor*), **Pre-Sub** (Predicate of *major*/Subject of *minor*), **Sub-Sub** (Subject of both), **Pre-Pre** (Predicate of both).

These concepts were first defined by *Aristotle* (Fait, 2023), from whom originated the first three forms, and later expanded upon by *Galen*, who added the fourth form (Łukasiewicz, 1957). In order to determine the validity of the normalized syllogisms, we used the *Table of Valid Syllogistic Forms* (Appendix F), which was defined in (Copi et al., 2018).

For the sake of this experiment, we tested the **Llama-3** (AI@Meta, 2024) and **gemini-3-flash-preview** (Google DeepMind, 2025) models. The smaller Llama models (tested both the 3B and 8B variants) struggle with term mapping especially when the terms take more complex forms[2].

The larger model shows significantly higher proficiency in term mapping, being capable of performing both simplification steps at the same time, which reduces computation time and resource consumption.

#### 3.3.2 Extending for Subtask 2

For the second subtask, we use the same model to simplify the sentences as in the first subtask, but we extend the algorithm that verifies validity. To handle multiple premises, we map the relations in a graph search problem. Every sentence form can be mapped to a relation in the following way: **A**⇒ *inclusion*, **E**⇒ *disjunction*, **I**⇒ *overlap*, **O**⇒ *non-inclusion*. These relations exhibit different structural properties. *Overlap* and *disjunction* are **symmetric** relations, whereas *inclusion* and *non-inclusion* are **asymmetric** ($A \subseteq B$ does not imply $B \subseteq A$). Both the *overlap/disjunction* pair and the *inclusion/non-inclusion* pair are contradictory. Among these, only inclusion has transitive properties. Because of this, relations propagate through

[2]In the case of *"Every cat is an invisible creature [...] a portion of animals are invisible,"* the model was not capable of correlating *"invisible creature"* with *"invisible"*.

## 4 Experimental Results and Discussion

### 4.1 Subtask 1

| Metric | Baseline LoRA | Augmented LoRA | Final Ensemble |
|---|---|---|---|
| Acc | 0.9319 | 0.9476 | **0.9686** |
| Intra-Plaus. CE | 0.0957 | **0.0426** | 0.0636 |
| Cross-Plaus. CE | 0.0749 | **0.0426** | 0.0428 |
| Total CE | 0.0853 | **0.0426** | 0.0532 |
| Acc/CE Ratio | 10.9218 | **22.2696** | 18.2094 |
| **Quadrant Acc.** | | | |
| Valid-Plaus. | 0.9792 | 0.9792 | **1.0000** |
| Valid-Implaus. | 0.9792 | 0.9375 | **1.0000** |
| Invalid-Plaus. | 0.8085 | **0.9149** | 0.8936 |
| Invalid-Implaus. | 0.9583 | 0.9583 | **0.9792** |

Table 1: Post-evaluation Neural Validity model performance metrics on the Subtask 1 test set.

**Neural Validity model results:** Using a baseline LLM, we observed that fine-tuning significantly improved model accuracy on the validation split, increasing it from around $64\%$ with direct prompting methods to approximately $93\%$. However, error analysis revealed that even after fine-tuning, the model remained vulnerable to content bias. During the evaluation phase, the official test results achieved ***accuracy:*** $92.15\%$ ***and a content bias of*** $7.47$. Transitioning from a standard fine-tuning regime to our proposed ensemble framework improved logical accuracy and the robustness-to-bias ratio. Further, integrating the **Contrastive Augmented Curriculum** and the **Logit Margin Calibration** ($\tau = 0.520$), we further neutralized these residual biases. Adding these modifications we received a significant increase in accuracy ($92.15\%$ $\Rightarrow 96.86\%$) and a decrease in logical bias ($7.47 \Rightarrow 4.26$) in post-evaluation. Analyzing the errors produced by this modified model (see Appendix A), we observed that all remaining errors **are false positives** and, with the exception of one case, nearly all **are influenced by plausibility bias**. Consequently, the model exhibits a residual tendency to favor semantically plausible conclusions. Moreover, our results reveal a trade-off between bias neutralization and raw performance across the different experimental stages. While the **Augmented LoRA** established a peak Acc/CE ratio of **22.27** and the lowest total content effect (**0.0426**), the **Final Ensemble** prioritized global logic, achieving a maximum accuracy of **0.9686**. This suggests that contrastive augmentation is the primary mechanism for bias reduction, while the ensemble smooths structural errors by aggregating over independent decision boundaries, diluting the curriculum's tar-

geted effect but improving overall accuracy.

**Neuro-Symbolic Logical Solver results:** The simplified logical model achieved strong results. By design, this model should not be vulnerable to content bias, so we examined the training errors more closely (see Appendix B)[3]. From an initial inspection of the errors, we observed that none were caused by sentence simplification. Instead, all of them appear to originate from the training data. Out of the eleven errors in train, eight are false negatives, and many of them seem to be susceptible to plausibility bias. There is only one false positive that does not follow plausibility bias [4]. This particular instance exhibits a contradiction, where the first premise denies that a square is a shape, whereas the second affirms that some shapes are squares. The conclusion's validity is therefore contingent on which premise is assumed to be true.

Our submission achieved an official result of $98.43\%$ *accuracy with* $3.19$ *content bias*, placing us in 22nd place. Post-evaluation we analyzed the errors in these results (see Appendix C) and similar to the training data, it appears that the errors originate from the ground truth labels. This would suggest that the true accuracy of our model should be $100\%$. In the official evaluation, we also obtained $97.91\%$ *accuracy with* $2.13$ *content bias*, placing us 17th on the leaderboard. However, the post-evaluation analysis revealed that this model is actually worse, its superior results being a product of shared biases with the test data. All three errors in the test data are false negatives that do not align with plausibility bias, which significantly skews our content bias score. Notably, two of these instances are structurally identical: the first premise is irrelevant to the conclusion, and both follow the form *All A are B* $\Rightarrow$ *Some B are A*, which should be true, but the test data labels it as false.

### 4.2 Subtask 2

In this subtask, we encountered several edge cases resulting from our implementation of the algorithm being incomplete (see Appendix G). Our official result on the evaluation phase was $97.37\%$ *accuracy* $9.47\%$ *F1 for premise retrieval and* $3.12$ *content bias* (15th place). In post-evaluation analysis, we traced the low premise retrieval score to an indexing error. Solving this has given us perfect premise retrieval on our valid results. A diagnostic evaluation of the test errors (see Appendix D) revealed that three of the five recorded failures stemmed from algorithmic and implementation limitations. By resolving an implementation error that accounted for two of these wrong predictions, our post-evaluation results reached *an accuracy of* $98.42\%$*, a premise F1 score of* $98.95\%$*, and a content bias of* $1.11$. The remaining two errors are ground truth errors, both of which are false negatives involving only a relevant premise. Together with the conclusion, they take the following form: *All A are B* $\Rightarrow$ *Some A/B are B/A*, which should be considered valid statements. We observed that these errors in the test data are very similar to the one error found in the test data of the first subtask.

## 5 Conclusion and Future Work

In this work, we explored two complementary strategies for disentangling logical reasoning from content-driven bias in large language models. One is a fully LLM-centered approach that combines contrastive data augmentation, ensemble modeling, and logit margin calibration to reduce belief bias and improve accuracy. The other is a hybrid neuro-symbolic framework, which decomposes the task by using an LLM as a linguistic encoder to map syllogisms into symbolic logic, which is then solved deterministically. This approach proved highly reliable and resistant to linguistic ambiguity. Our analysis reveals that parts of the task annotations appear to have been susceptible to plausibility bias, leading to several recurring error patterns in the ground truth data.

Future work could further improve the deterministic reasoning component, potentially reducing the task to a pure sentence normalization problem. In addition, both approaches could be evaluated on more complex reasoning settings. For the neural approach, this includes handling longer reasoning chains and additional premises. For the hybrid framework, natural extensions include multilingual syllogistic reasoning and broader forms of symbolic inference. Finally, it is important to note that the dataset used in this task contains relatively clean syllogistic structures, where relations are expressed in short sentences between two groups. A valuable direction for future research is therefore to evaluate both approaches on more realistic and linguistically complex reasoning scenarios.

[3] As the model undergoes no task-specific training, the full training set was repurposed for evaluation.

[4] *"A square cannot be a shape. There exist shapes that are squares and are also rectangles. It necessarily follows that a subset of rectangles are not shapes."*

## References


AI@Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lorenzo 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)*.

Irving M. Copi, Carl Cohen, and Victor Rodych. 2018. *Introduction to Logic*, 15th edition. Routledge, New York. Also available as eTextbook (ISBN 9781351386975).

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: Human Language Technologies (NAACL)*.

Paolo Fait. 2023. *Aristotle and Theophrastus*, page 37–56. Cambridge Companions to Philosophy. Cambridge University Press.

Google DeepMind. 2025. Gemini 3: Frontier-scale modeling and reasoning capabilities. Technical report, Google.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

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

Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. 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 (IJCNLP-AACL)*.

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, Ryo Ando, Terufumi Morishita, Hiroki 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*.

L. Ranaldi, M. Valentino, and A. 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)*.

T. Seals and V. 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: Human Language Technologies (NAACL)*.

M. Valentino, G. Kim, D. Dalal, Z. Zhao, and A. 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)*.

M. Wysocka, D. Carvalho, O. Wysocki, M. Valentino, and A. Freitas. 2025. Syllobio-nli: Evaluating large language models on biomedical syllogistic reasoning. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

J. Xu, H. Fei, L. Pan, Q. Liu, M. Lee, and W. Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*.

Jan Łukasiewicz. 1957. *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic*, 2nd enlarged edition. Clarendon Press, Oxford.

## A Appendix: table of Neural Validity Experiments Results

| Syllogism | Plausibility | Ground Validity | Predicted Validity |
|---|---|---|---|
| The category of chairs and the category of living things do not overlap. The group of living things and the group of inanimate objects are mutually exclusive. A portion of inanimate objects are not chairs. | True | False | True |
| Some member of the group of fish is not in the group of mammals. At least one dog is a mammal. At least one dog is not a fish. | True | False | True |
| There is at least one spoon that is a kitchen tool. Some of the utensils are spoons. Consequently, some of the utensils are kitchen tools. | True | False | True |
| The category of celestial bodies and the category of planets are mutually exclusive. Not a single planet is a star. It must be the case that a portion of stars are not celestial bodies. | False | False | True |
| Not a single animal with wings is an aquatic mammal. Anything that is an aquatic mammal is a fish. It is necessarily true that no fish are animals with wings. | True | False | True |
| There is at least an odd number that is a negative number. A portion of the prime numbers are negative numbers. Consequently, some prime numbers are odd numbers. | True | False | True |

Table 2: Testing errors from our modified fine-tuned model

## B Appendix: table of training errors on Neuro-Symbolic model

| Syllogism | Plausibility | Ground Validity | Predicted Validity |
|---|---|---|---|
| Every single animal is a living thing. Anything that is a bird is an animal. All birds are living things. | True | False | True |
| Anything that is a dog is a mammal. There are no sharks which are also dogs. It is the case that some sharks are not mammals. | True | True | False |
| There are no animals that are also plants. At least one tree is a plant. At least one tree is not an animal. | True | False | True |
| Anything that is a spider is a mammal. All spiders are insects. This means that a few insects are mammals. | False | False | True |
| Every object that orbits a star is a planet. Anything that is a gas giant is an object that is a planet. Therefore, there are no gas giants that do not orbit a star. | True | True | False |
| Not a single person who is a poodle is a dog. Every single animal is a poodle. The conclusion that no animal is a dog is inescapable. | False | False | True |
| Every car is a motorized vehicle. At least one bicycle is not a motorized vehicle. Not all bicycles can be considered cars. | True | False | True |
| The group of furniture is a subset of the group of chairs. At least one sofa is not a chair. Not every sofa is a furntiture. | False | False | True |
| It is true that everything that is an animal is a mammal. It is also true that no reptiles are animals. Thus, it is concluded that no reptiles are animals. | False | False | True |
| A square cannot be a shape. There exist shapes that are squares and are also rectangles. It necessarily follows that a subset of rectangles are not shapes. | False | True | False |
| Every vehicle has wheels. At least one car does not have wheels. Not all cars are vehicles. | False | False | True |

Table 3: Training errors

## C Appendix: table of test errors from the Neuro-Symbolic model in Subtask 1

| Syllogism | Plausibility | Ground Validity | Predicted Validity |
|---|---|---|---|
| There are no bikes that can be called cars. It is also true that every bike is a type of vehicle. This has led to the conclusion that a portion of vehicles are bikes. | True | False | True |
| Something that is a chair is a table. A thing that is a table cannot be a building. It follows that a building is never a chair. | True | False | True |
| There are no cats that can be called dogs. It is also true that every cat is a type of animal. This has led to the conclusion that a portion of animals are cats. | True | False | True |

Table 4: Testing errors in subtask 1

## D Appendix: table of test errors from the Neuro-Symbolic model in Subtask 2

For readability the table below presents only the premises that were considered relevant.

| Syllogism | Plausibility | Ground Validity | Predicted Validity |
|---|---|---|---|
| It is also true that all carrots are vegetables. Therefore, it is concluded that a few vegetables are carrots. | True | False | True |
| Anything which is a guitar is also a string instrument. All string instruments are things considered musical instruments. We can infer that some musical instruments are guitars. | True | True | False |
| Every single person who is a carpenter is a builder. It is the case that every cabinet maker is a carpenter. A certain number of cabinet makers are builders. | True | True | False |
| Any item that is a vegetable is also a plant. There are some foods that are green and are also vegetables. Consequently, some green foods are plants. | True | True | False |
| All of the things that are windows are planets. It can be deduced that there are windows that are planets. | False | False | True |

Table 5: Testing errors in subtask 2(relevant premises)

## E Appendix: PCA Geometric Comparison

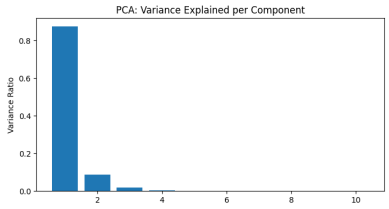


Figure 4: Baseline: Entangled

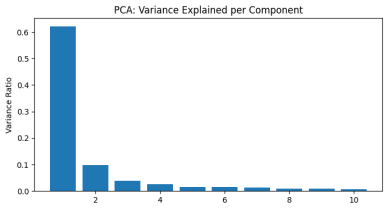


Figure 5: Augmented: Decoupled

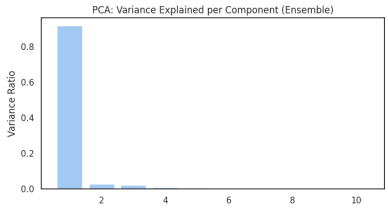


Figure 6: Ensemble: Exploded

Figure 7: Comparative interpretation of Figures 4–6, illustrating the PCA projections across the three experimental stages. The transition from (left) dense semantic clustering to (right) a more distributed latent structure reflects a reduced dominance of plausibility-aligned variance.

In this section, we provide visual evidence to support the representational trends described in results section. Principal Component Analysis (PCA) was applied to hidden-state representations extracted from each experimental stage in order to examine how the variance structure evolves under successive interventions.

The three figures illustrate changes in the geometric relationship between logical validity and semantic plausibility across training stages.

**Stage 1: Baseline Entanglement.** In the baseline LoRA model, PCA projections reveal a concentrated variance structure in which logical validity and semantic plausibility are not geometrically well separated. The dominant principal component captures a large proportion of total variance, suggesting that semantic plausibility may act as a primary organizing factor in the latent space. This concentration may contribute to the model's vulnerability to belief-bias conflicts.

**Stage 2: Augmented Reduction.** Following the introduction of the Contrastive Augmented Curriculum, the PCA manifold exhibits increased dispersion and a reduced dominance of plausibility-aligned variance components. The variance becomes more distributed across principal components, indicating a partial decoupling between structural validity and semantic believability. This redistribution aligns with the observed reduction in Total Content Effect.

**Stage 3: Manifold Expansion (Majority Vote Ensemble).** In the final ensemble configuration, the latent representation becomes more broadly distributed. Because the majority vote aggregates independent predictions from separately fine-tuned adapters, the resulting decision space is less dependent on a single model's internal bias structure. The expanded variance profile suggests a further weakening of a dominant semantic axis, consistent with the ensemble's improved logical accuracy and robustness.

Overall, the PCA analysis indicates that successive training interventions progressively redistribute representational variance away from plausibility-dominated directions and toward more structurally differentiated manifolds. While PCA does not directly identify semantic features, the geometric trends are consistent with the quantitative bias-reduction metrics reported.

## F Appendix: Table of Valid Syllogistic Forms

Table 6: Valid Syllogistic Moods by Figure

| Mood | Fig. 1 | Fig. 2 | Fig. 3 | Fig. 4 |
|---|---|---|---|---|
| **AAA** | X | | | |
| **AAI** | X* | | X | X |
| **AEE** | | X | | X |
| **AEO** | | X* | | X* |
| **AII** | X | | X | |
| **AOO** | | X | | |
| **EAE** | X | X | | |
| **EAO** | X* | X* | X | X |
| **EIO** | X | X | X | X |
| **IAI** | | | X | X |
| **OAO** | | | X | |

*Note: Moods marked with X* denote "subaltern" or weakened forms that rely on the assumption that the subject set is not empty (Existential Import).

## G Appendix: Explaining algorithm shortcomings

We distinguish between the inherent limitations of our current algorithmic design and specific implementation errors identified during post-evaluation.

### G.1 Algorithmic Limitations

Due to the current scope of implementation, our inclusion graph is treated as a **doubly linked list** rather than a branching tree or directed acyclic graph. This architecture creates a specific **edge case**: in instances involving multiple direct inclusions for a single term, subsequent relations overwrite previous ones. This prevents the system from maintaining a full hierarchy of logical containment.

### G.2 Implementation Errors

Further analysis revealed two primary implementation errors that impacted our final metrics:

- **Indexing Discrepancy**: A misalignment in the indexing logic caused a significant drop in our premise retrieval $F_1$ score. This was a procedural error rather than a failure of the model's logical reasoning.
- **Overlap Logic**: We identified a missing condition in the inference engine where values within the same inclusion graph were not correctly flagged as overlapping. This led to missed correlations in specific syllogistic moods.