Ellat at SemEval-2026 Task 11: Comparing Encoder and Decoder Models for Syllogistic Reasoning

Farzaneh Bayan Memar Hanneke Huls Matthijs ten Hove
University of Groningen
{f.bayan.memar, j.s.huls, m.h.ten.hove}@student.rug.nl

Abstract

For SemEval-2026 Task 11 (Subtask 1: English), Team Ellat investigates whether language models can assess logical validity independently of semantic plausibility. Since these models learn statistical patterns instead of explicit logical rules, they often rely on world knowledge and semantic shortcuts rather than formal logic. To address this challenge, we evaluate three architectures: MiniLM-L6-mnli-binary, DeBERTa-v3-small, and Llama 3.1-8B-Instruct, applying task-specific fine-tuning for encoder models and Abstract Logic Augmentation with QLoRA for LLaMA. DeBERTa achieved the strongest overall performance, MiniLM showed clear reductions in content bias after fine-tuning, and Llama 3.1-8B exhibited strong plausibility bias in the zero-shot setting. However, our augmented fine-tuning approach led to only modest improvements and a partial shift toward structure-based reasoning. Overall, fine-tuning and abstraction-based augmentation reduce plausibility bias, but fully separating logical validity from semantic content remains challenging across architectures.

1 Introduction

SemEval-2026 Task 11 (Subtask 1) (Valentino et al., 2026b) investigates whether language models can judge logical validity independently from semantic plausibility. We participated using the English dataset. The task is based on the idea that in Large Language Models, meaning and logical structure are often strongly intertwined. Since these models learn statistical patterns instead of explicit logical rules, they may rely on word associations and world knowledge rather than formal logic. The dataset contains syllogisms annotated on two independent binary dimensions: logical validity (true/false) and plausibility (true/false). This allows analysis of the entanglement between meaning and structure. This is important for real-world applications, where models must judge arguments correctly even if meaning and logic conflict. Our strategy is to compare models with different capacities and training strategies and to test methods that strengthen logical reasoning. We evaluate MiniLM-L6-mnli-binary (Wang et al., 2020), DeBERTa-v3-small (He et al., 2021), and Llama 3.1-8B-Instruct (Dubey et al., 2024). For the encoder models, we apply task-specific fine-tuning to reduce plausibility bias. For LLaMA, we introduce *Abstract Logic Augmentation*, replacing content words with symbolic placeholders while keeping the logical form. This reduces semantic and world-knowledge cues. We then fine-tune LLaMA with QLoRA (Dettmers et al., 2024) to see if abstraction-focused training improves validity predictions. Our results show that fine-tuning reduces content bias in smaller models, but plausibility effects are still difficult to remove completely. On the official leaderboard, DeBERTa-v3-small ranked 33rd out of 45 teams (34.22 score), MiniLM-L6 ranked 39th, and our fine-tuned LLaMA models ranked 45th. Although the rankings are modest, our analysis shows that Abstract Logic Augmentation and QLoRA give small but meaningful improvements. Overall, separating logical structure from semantic content remains challenging across architectures.

2 Background

2.1 Task Description

We participated in Subtask 1 (English) of SemEval-2026 Task 11 (Valentino et al., 2026b). The task evaluates whether language models can assess *logical validity* independently of *semantic plausibility*. Each instance in the set consists of a syllogistic argument composed of two premises and a conclusion. Every syllogism is annotated along two independent dimensions:

- **Logical validity:** whether the conclusion logically follows from the premises (true/false)

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2130–2138

- **Plausibility:** whether the argument aligns with real-world knowledge (true/false)

This results in four possible combinations: valid/plausible, valid/implausible, invalid/plausible, and invalid/implausible. In Subtask 1, systems are required to predict logical validity only. The plausibility label is provided to enable evaluation of content-driven bias.

The most informative instances in the dataset are *conflict cases*, where logical validity and real-world plausibility disagree. These cases make it possible to determine whether a model relies on semantic world knowledge or on formal logical structure.

We provide two illustrative examples from the training set:

Valid but Implausible:

- *Premise 1:* Plants are in no way living organisms.
- *Premise 2:* All things that are trees are living organisms.
- *Conclusion:* Trees cannot be classified as plants.

Although the premises contradict real-world knowledge, the conclusion follows logically from the premises.

Invalid but Plausible:

- *Premise 1:* There are some animals that are mammals.
- *Premise 2:* Something that is a dog is an animal.
- *Conclusion:* Consequently, every dog is a mammal.

The conclusion is factually true in the real world, but it does not logically follow from the premises. The organizers provide predefined training, development, and test splits. All instances in the dataset are written in English and follow controlled syllogistic structures, enabling systematic analysis of reasoning behavior.

2.2 Related Work

The task builds on a growing body of research examining whether large language models perform genuine logical reasoning or rely on semantic cues. Prior studies have shown that LLMs are susceptible to the *content effect*, where logical judgments are influenced by real-world plausibility rather than formal validity. Dasgupta et al. (2022) demonstrate that models exhibit human-like belief bias in reasoning tasks. Kim et al. (2025) further analyze internal reasoning patterns in syllogistic evaluation, while Valentino et al. (2026a) propose strategies to reduce content-driven bias.

Other work has evaluated reasoning in alternative strategies. Seals and Shalin (2024) assess LLMs using the Wason selection task and find that models perform better when arguments contain familiar semantic content, suggesting reliance on learned associations. Saparov and He (2023) argue that LLMs often do not fully analyze the logical structure. Instead, they make a fast decision based on patterns they have seen often before. Similarly, Itzhak et al. (2024) show that instruction tuning can strengthen belief bias, leading models to select plausible conclusions even if they are logically incorrect.

Our approach connects to recent work that attempts to encourage structure-based reasoning by abstracting away semantic content. Ranaldi et al. (2025) demonstrate that replacing real words with symbolic placeholders can improve logical generalization. Inspired by this line of research, we apply *Abstract Logic Augmentation*, replacing content-specific terms with symbols in order to reduce the availability of semantic shortcuts while preserving formal structure.

3 System Overview

To investigate whether language models can distinguish logical validity from semantic plausibility, we implemented and evaluated three systems with different architectures and training strategies: (1) MiniLM-L6-mnli-binary, (2) DeBERTa-v3-small, and (3) Llama 3.1-8B-Instruct. Our design allows us to compare a compact NLI-based classifier, a lightweight encoder model, and a large instruction-tuned LLM under controlled conditions. Figure 1 provides an overview of the three system architectures.

Across all systems, our primary objective is not only to optimize predictive performance, but to assess whether predictions are grounded in abstract logical structure rather than semantic plausibility cues.

3.1 Overall Pipeline

For all models, the input consists of two premises and a conclusion forming a syllogistic argument. The target label is *logical validity* (true/false), while plausibility labels are only used for analysis.

- **Abstract Logic Augmentation:** Replace content-specific nouns with symbolic placeholders while preserving logical structure for our LLaMA model.
- **Model inference or fine-tuning:** Train or evaluate the model on validity classification.
- **Bias analysis:** Analyze predictions across validity–plausibility combinations to measure content sensitivity

The three systems differ primarily in their training strategy and degree of abstraction-oriented intervention.

3.2 MiniLM-L6-mnli-binary

We used the MiniLM-L6-mnli-binary model (Wang et al., 2020) in two configurations: (1) as a frozen pre-trained baseline and (2) after task-specific fine-tuning.[1] MiniLM-L6 is designed for binary Natural Language Inference (NLI), predicting entailment vs. not_entailment. We reformulate syllogistic validity as an entailment task by treating the premises as input text and the conclusion as the hypothesis, labeling a syllogism as valid if the conclusion is entailed. The frozen model measures how well general NLI transfers to logical reasoning, while the fine-tuned version is trained on the task data using cross-entropy loss to test whether supervised adaptation reduces plausibility-driven errors. To assess content bias, we compare both configurations across all validity–plausibility combinations, examining whether fine-tuning shifts predictions toward structure-based reasoning rather than semantic cues.

3.3 DeBERTa-v3-small

We additionally evaluated Microsoft DeBERTa-v3-small (He et al., 2021)[2], a lightweight encoder architecture designed for efficient natural language understanding.

The model receives the syllogism as input and outputs a binary prediction (valid/invalid). Fine-tuning was performed using supervised training on the official training split.

[1] Hugging Face model card
[2] Hugging Face model card

3.4 Llama 3.1-8B-Instruct

Our third system is `meta-llama/Meta-Llama-3.1-8B-Instruct` (Dubey et al., 2024), a large instruction-tuned language model. Preliminary zero-shot evaluation revealed a strong prediction bias, with the model frequently defaulting to an "invalid" label. This behavior suggests a reliance on heuristic decision strategies rather than formal logical reasoning. To address this, we implemented two primary interventions:

Abstract Logic Augmentation. We transform instances by replacing specific category terms (e.g., *animals, buildings*) with single-letter symbolic placeholders (e.g., *A, B, C*) to minimize semantic interference. The mapping is consistent within each syllogism to preserve logical relations (e.g., *"All A are B. Some B are C. Therefore, some A are C."*) This transformation aims to remove real-world knowledge cues and reduce the availability of semantic shortcuts. This augmentation was applied as a preprocessing step before fine-tuning, such that the model was trained on abstracted syllogisms.

Parameter-Efficient Fine-Tuning (QLoRA). After augmentation, we fine-tuned the model using QLoRA with 4-bit quantization for 60 steps (Dettmers et al., 2024), utilizing the Unsloth library for memory efficiency. The training prompt was specifically designed to encourage the model to prioritize logical structure over factual familiarity. By evaluating this configuration, we assess whether targeted abstraction-oriented training can shift LLM predictions toward formal validity.

4 Experimental Setup

4.1 Preprocessing

We used the official train/development/test splits provided by the task organizers. All data used in this work is in English. The development set was used for model selection and analysis; final results are reported on the test set via the leaderboard.

In the MiniLM NLI setup, premises form the premise and the conclusion the hypothesis. DeBERTa uses the full syllogism as a single sequence. For LLaMA with Abstract Logic Augmentation, nouns were replaced by single-letter symbolic placeholders (e.g., A, B, C) with consistent mappings per syllogism.

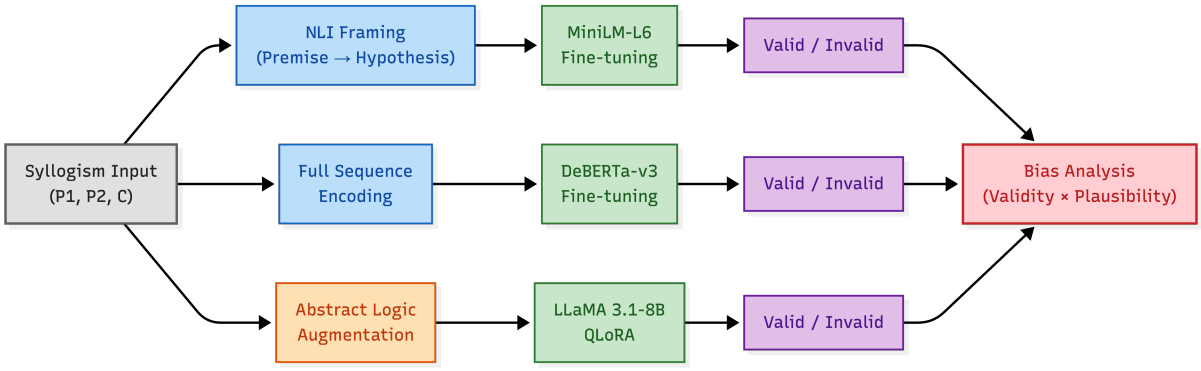

Figure 1: Overview of the three system architectures for syllogistic validity classification.

4.2 Training Configurations

All encoder models were trained using supervised cross-entropy loss for binary classification.

MiniLM-L6-mnli-binary. Experiments were run in Google Colab (Python 3, HuggingFace `transformers`). The model was fine-tuned for 3 epochs (batch size 16, learning rate 2×10^{-5}, weight decay 0.01).

DeBERTa-v3-small. Experiments were conducted on Kaggle (NVIDIA T4). The model was fine-tuned for 4 epochs (batch size 8, learning rate 3×10^{-5}) without weight decay.

Llama 3.1-8B-Instruct. Experiments were run on the Hábrók high-performance computing cluster (NVIDIA A100 GPU). Zero-shot evaluation used no parameter updates. For fine-tuning, we applied QLoRA with 4-bit quantization (Unsloth library) and trained for 60 steps (learning rate 3×10^{-4}).

4.3 Evaluation Metrics

The official evaluation includes:

- **Validity Accuracy (%)**: percentage of correct validity predictions across all items.
- **Total Content Effect (TCE)**: a composite score measuring the average accuracy difference due to plausibility across all four validity–plausibility conditions. A lower TCE indicates less content bias.
- **Combined Score**: the primary ranking metric, defined as $\frac{\text{ACC}}{1+\ln(1+\text{TCE})}$, rewarding high accuracy while penalizing content bias.

During development, we additionally report class-wise F1 scores to assess performance symmetry across valid and invalid classes.

Model	F1 valid	F1 invalid
MiniLM-L6	0.82	0.81
DeBERTa-v3-small	0.85	0.85
Llama 3.1-8B (zero-shot)	0.20	0.64

Table 1: F1 scores on the development set.

5 Results

5.1 Development Set Performance

During development, we evaluated all models on the official validation set to determine which systems to submit. Table 1 shows the F1 scores on the development set.

DeBERTa-v3-small achieved the most balanced performance (F1 = 0.85 for both classes). MiniLM-L6 also performed strongly with symmetric results (0.82 / 0.81). In contrast, Llama 3.1-8B in zero-shot mode exhibited a strong negative prediction bias, correctly identifying only 13% of valid arguments (F1 = 0.20 for the valid class). Rather than systematic logical reasoning, the model appeared to rely on simple heuristics. Despite its low performance, we decided to fine-tune LLaMA to investigate whether Abstract Logic Augmentation could reduce this bias.

5.2 Official Leaderboard Performance

Table 2 presents the final leaderboard results on the test set.

Our best system, DeBERTa-v3-small, ranked 33rd out of 45 teams. It achieved the highest validity accuracy and the lowest content effect (3.12), resulting in our best combined score (34.22). MiniLM-L6 ranked 39th, showing moderate performance but higher content sensitivity. The Llama 3.1-8B submissions ranked 45th. The improved fine-tuned version achieved small gains over the ini-

Submission	Validity Acc. (%)	TCE	Combined Score
MiniLM-L6 Fine-tuned	74.35	9.73	22.04
DeBERTa Fine-tuned	82.72	3.12	34.22
LLaMA Initial Fine-tuned	50.26	50.00	10.19
LLaMA Improved Fine-tuned	53.93	47.92	11.03

Table 2: Final leaderboard results (test set).

	Plausibility	
Validity	TRUE	FALSE
FALSE	41	38
TRUE	12	29

Table 3: Misclassifications baseline MiniLM (dev-set)

	Plausibility	
Validity	TRUE	FALSE
FALSE	15	12
TRUE	8	8

Table 4: Misclassifications fine-tuned MiniLM (dev-set)

tial submission, reducing the content effect (50.00 $\rightarrow$ 47.92). Although performance remained below encoder-based models, this reduction suggests that Abstract Logic Augmentation partially mitigates plausibility bias.

6 Analysis

6.1 DeBERTa

Although DeBERTa achieved the strongest overall performance, we examined whether its predictions were influenced by plausibility cues.

On the development set, it produced 29 additional correct predictions (12.1%) across 240 conflict cases. These improvements were evenly distributed across syllogism types (some, all, none) and across all validity–plausibility combinations, rather than concentrated in a specific logical form.

This pattern suggests that DeBERTa does not rely on plausibility cues but applies its learned decision boundaries consistently across structural variations.

6.2 MiniLM-L6: Effect of Fine-Tuning

To examine content bias, we analyzed misclassifications on the development set.

Baseline (Frozen) Model Table 3 shows the misclassification distribution for the frozen baseline model. The baseline made more errors in valid but implausible cases (29) than in valid and plausible cases (12), indicating reliance on semantic plausibility cues.

Fine-Tuned Model Table 4 shows the misclassification distribution after fine-tuning.

After fine-tuning, errors were balanced across plausibility conditions, suggesting that supervision substantially reduced content bias.

6.3 Llama 3.1-8B: Effect of Fine-Tuning

Zero-Shot Baseline The zero-shot LLaMA model achieved 50.1% accuracy but showed a strong negative bias, predicting "invalid" for most inputs (valid recall = 13%).

Fine-Tuned Models Fine-tuning corrected 70 previously misclassified syllogisms: 35 belief-bias cases, 24 semantic-implausibility cases and 11 mixed cases. This shows that instruction-tuned LLMs rely heavily on world knowledge, while Abstract Logic Augmentation encourages greater focus on formal structure.

Example (Semantic Implausibility)
Example 1:

- *Premise 1:* Everything that is a car is a sentient cloud.
- *Premise 2:* All sentient clouds are plants.
- *Conclusion:* Some plants are cars.

Although absurd, the argument is logically valid. The zero-shot model was misled by content, whereas the fine-tuned model recognized the correct structure. To further assess the impact of fine-tuning and Abstract Logic Augmentation, we analyzed cases of *belief bias*—syllogisms that are structurally invalid but have highly plausible conclusions. We identified 18 such instances in the evaluation set where baseline models failed but fine-tuned models succeeded (see Appendix A).

Example 2:

- *Premise 1:* Anything that is a bird is a vertebrate.
- *Premise 2:* Every bat is a vertebrate.
- *Conclusion:* Nothing that is a bat is a bird.

Although the conclusion is factually true, the argument is logically invalid. The baseline MNLI and zero-shot Llama-3.1-8B models were misled by real-world knowledge and predicted it as valid. After fine-tuning, LLaMA correctly classified it as invalid, consistent with DeBERTa. This indicates that fine-tuning helps models suppress factual plausibility in favor of formal structure.

6.4 Robustness to Semantic Implausibility

We also examined cases where semantic implausibility conflicted with logical validity. We found 8 instances where all models except fine-tuned DeBERTa failed (see Appendix B).

For example:

- *Premise 1:* Every planet is made of cheese.
- *Premise 2:* The Earth is a planet.
- *Conclusion:* Part of Earth is made of cheese.

Despite being logically valid, the argument is semantically absurd. Three models relied on world knowledge and predicted it as invalid. DeBERTa, however, correctly identified the valid structure, demonstrating a stronger separation between formal reasoning and factual knowledge.

Limitations

Due to resource constraints, we could not conduct a full ablation study of QLoRA and Abstract Logic Augmentation. Although zero-shot LLaMA achieved similar overall accuracy (50.00%) to the fine-tuned model (50.26%), this hides severe class imbalance (valid $F1 = 0.20$). Improvements are better captured by class-wise F1 and Content Effect than by overall accuracy.

7 Conclusion

We examined whether language models can distinguish logical validity from semantic plausibility in SemEval-2026 Task 11 (English Subtask 1), comparing MiniLM-L6, DeBERTa-v3-small, and Llama 3.1-8B. Fine-tuning reduced plausibility bias in encoder models. DeBERTa achieved the strongest and most robust performance, while MiniLM improved substantially over its frozen baseline. Zero-shot LLaMA relied heavily on plausibility. Abstract Logic Augmentation with QLoRA reduced content bias and improved logical consistency, but performance remained below encoder models. Separating logical structure from world knowledge remains difficult for large instruction-tuned LLMs. Future work may explore larger datasets, multi-task learning, or larger models (e.g., Llama 3.1-70B), as well as structured prompting and explicit reasoning supervision.

Ethical Considerations

LLMs often favor semantic plausibility over formal logic. In high-stakes domains (e.g., legal analysis or fact-checking), this may produce convincing but logically incorrect conclusions. Careful human oversight is therefore essential.

Acknowledgments

We are grateful to Malvina Nissim and Huiyuan Lai for their valuable guidance, feedback, and support throughout this project at the University of Groningen. LLaMA experiments were conducted using the Hábrók high-performance computing cluster provided by the University of Groningen.

References

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*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.22706*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to Bias:

Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. *Transactions of the Association for Computational Linguistics*, 12:352–371.

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*. Association for Computational Linguistics.

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

Abulhair Saparov and He He. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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 (NAACL)*.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026a. Mitigating content effects on reasoning in language models through fine-grained activation steering.

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026b. 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)*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

A Belief Bias Examples

Table 5 presents the complete list of 18 syllogisms from our evaluation set that exhibit belief bias. In these specific cases, the argument is logically invalid (*Validity = False*), but the conclusion aligns perfectly with real-world facts (*Plausibility = True*).

Because of this semantic plausibility, both the baseline MNLI and the zero-shot Llama 3.1-8B models were misled and incorrectly predicted these instances as valid. However, our fine-tuned LLaMA and DeBERTa models successfully ignored the factual distractor and correctly identified the logical fallacy in all 18 cases.

B Implausible Cases Solved Exclusively by DeBERTa

Table 6 lists the 8 specific instances from the evaluation set where the fine-tuned DeBERTa model successfully predicted the correct label, while all three other tested models (MNLI, zero-shot LLaMA, and fine-tuned LLaMA) failed. These cases are characterized by high semantic absurdity (implausibility), which misled the other models into making incorrect predictions despite the underlying logical structure.

Dataset ID	Syllogism (Invalid Logic, Plausible Conclusion)
68b29c80-27f6-4121-8408-295bd8de466c	Anything that is a bird is a vertebrate. It is also the case that every single bat is a vertebrate. Consequently, nothing that is a bat is a bird.
ee8e00e2-7bf8-41a3-8009-40d328cb1d60	It is true that no mammals are insects. There are no butterflies that are mammals. Therefore, it follows that some butterflies are insects.
312d903d-c5fc-4ac8-beb2-3ec72f6b25b9	There are no plants that are fish. Some objects classified as a rose is a plant. Therefore, no roses are fish.
65aaed4e-522b-422d-b4e4-cd5dfced222e	Anything that is a mammal is a human. Every dog is a mammal. not a single dog is a human.
95637ce9-5489-45cb-aa6c-73934ab7e489	Every single dog is an animal. Everything that is an animal is a mammal. It must therefore be true that every mammal is an animal.
6cb41e5e-32e3-4eeb-b3b2-700c7e3a692f	Anything that is a dog is an animal. Everything that is a cat is an animal. It is the case that some cats are not canines.
82567966-505e-45fe-a4b0-c34de685c0a7	Every single dog is a mammal. A small number of mammals are animals. Some animals are not dogs.
ec60f858-72d6-455b-976f-fe289c371eaf	Every dog is a mammal. no dog is a reptile. Therefore, no reptiles are mammals.
06a0aa7f-5aaa-4a9a-af0e-7cfc8affd23a	Everything that is a bird is an animal. Every single eagle is an animal. It is the case that some eagles are birds.
d8e827ad-b91a-45a7-8d28-7527f8b59d1d	nothing that is a plant is a mammal. A portion of trees are to be found among the plants. Thus, there are no trees that are also mammals.
7f71a082-108e-4c3c-a1ad-1af6c5bc3088	It is true that some people who are students are human. There are teenagers that are also students. This proves that a number of humans are not teenagers.
a943b701-942e-4df8-bb91-4cee9349502a	Some plants are a type of vegetable. There are some carrots that are vegetables. This leads to the conclusion that some carrots are plants.
c38a9bc0-efdd-49af-af59-17a6567d8817	Every item that is a mammal is an animal. Some lions are mammals. Therefore, all lions can be said to be animals.
166f5a4c-b0e7-4f2b-b25d-1c0a08bd59bb	Every single human being is a mammal. It is also true that every lion is a mammal. This means that there are no lions that are humans.
dc67f615-0f81-4c0b-8870-1a0a7d995ab6	Anything that is a bird is a vertebrate. There are no insects which are birds. Therefore, it can be concluded that no insects are vertebrates.
eff43e93-d2d9-4b07-877e-d63a8bdf260d	Every poodle is a dog. no cat is a poodle. Therefore, no cat is a dog.
a7d18367-6870-4eee-9a55-60a5dcc9dd70	Anything that is a dog is also a pet. There are some cats that are not pets. It must be true that no cat is a dog.
97bd1911-5aca-41a1-ba6f-835f2ba45abf	Every cat is a mammal. The set of dolphins contains no cats. Every single dolphin is a mammal.

Table 5: Complete list of 18 belief bias examples (Invalid but Plausible) successfully overcome by the fine-tuned models.

Dataset ID	Syllogism (Highly Implausible Semantic Context)
d8954215-16e1-4b1a-ab2f-bb98b633712d	There are some ideas that are made of cheese. Every single idea is an emotion. Consequently, some emotions are made of cheese.
0700415b-1bd2-4bf2-baba-848f70b38c7d	Some trains are not celestial bodies. It is the case that every single train is a planet. Consequently, some planets cannot be considered celestial bodies.
dc5e8a17-f9ee-4f1c-a6e1-e55e3893609d	All things which are ghosts are scientific facts. All things which are ghosts are mythical beings. It follows that some mythical beings are scientific facts.
941ec598-d637-4f22-b558-367267161a64	Every single planet is a thing made of cheese. The Earth is a planet. A certain part of Earth is made of cheese.
8be74042-6920-42ad-8f7f-3c0e53ba5fe7	Plants are never stones. A few stones are, in fact, flowers. There exist flowers that are not plants.
c2029414-d28e-4fbd-acc8-12f855acf46c	Every fictional characters is a stone. There are fictional characters which are animals. Thus, a portion of animals are stones.
176e3ca1-2c5c-4500-8b42-792ce58d4098	Everything that is a fish is also an insect. Everything that is a fish is also made of fire. Hence, some things that are made of fire are insects.
732e20b3-c504-4115-914d-496d0c4ee6da	Every single mammal is a living thing. The set of animals contains no living things. not a single mammal is an animal.

Table 6: Complete list of 8 highly implausible syllogisms where only the DeBERTa model predicted correctly.