SpyComet at SemEval-2026 Task 11: When Adversarial Debiasing Backfires - A Comparison of Data-Level and Model-Level Debiasing

C. Sai Aravind **Sunil Saumya** **C. Pothan Sai Reddy**
IIIT Dharwad
c.saiaravind@gmail.com, sunil.saumya@iiitdwd.ac.in,
cpothansai2006@gmail.com

Abstract

We describe our system for SemEval-2026 Task 11 Subtask 1, which requires classifying natural-language syllogisms as valid or invalid while minimizing the influence of content plausibility on predictions. Our system, **MLA-CI** (Multi-Layer Adversarial for Content Invariance), is a DeBERTa-v3-base classifier that combines multi-layer feature extraction, gradient-reversal adversarial training, structure-preserving template augmentation, implausible-class oversampling, and focal loss. On the official test set, MLA-CI achieves 79.06% accuracy with 4.17% content effect (combined score 29.92, 35th of 45 teams). Through systematic ablation on held-out validation data, we find that *adversarial training at standard strength is counterproductive* when template augmentation is present. Across three random seeds, removing adversarial training yields a mean combined score of 38.15 ± 5.32, compared to 26.41 ± 0.99 for the full system - with the worst ablated run still outperforming the best full-system run by 5.9 points. A sweep over seven adversarial pressure (λ) values confirms that only very light adversarial pressure ($\lambda \leq 0.1$) preserves accuracy, while the submitted strength ($\lambda = 1.0$) and above cause severe degradation. Our analysis reveals that gradient reversal over-suppresses plausibility-correlated features, disproportionately harming accuracy on plausible-valid syllogisms, and that data-level augmentation is a more effective and more stable debiasing strategy than model-level adversarial training for this task. Code is available at https://github.com/comet-web/semeval2026-task11-spycomet.[1]

1 Introduction

Syllogistic reasoning determining whether a conclusion follows from a pair of premises is a foundational task in formal logic. A syllogism consists of two premises and a conclusion, and is *valid* if the conclusion follows necessarily from the premises, regardless of whether the statements are true in the real world. Table 1 illustrates this distinction: the conclusion sounds plausible given everyday knowledge about cities and capitals, yet the argument is logically *invalid* because the conclusion does not follow from the premise structure.[2]

Component	Text
Premise 1	Every city is a location.
Premise 2	Anything that is a location is a capital city.
Conclusion	Therefore, every capital city is a city.
Validity	**Invalid**
Plausibility	**Plausible**

Table 1: An example syllogism from the SemEval-2026 Task 11 training set. Although the conclusion sounds believable, the logical structure is invalid—illustrating the content effect, where surface plausibility misleads both humans and models.

This mismatch between plausibility and validity is precisely what makes the task challenging. Both humans and language models exhibit *content effects*: the tendency to judge plausible-sounding conclusions as logically valid and implausible ones as invalid, conflating real-world believability with formal logical structure (Dasgupta et al., 2022; Eisape et al., 2024). SemEval-2026 Task 11 (Valentino et al., 2026) targets this problem directly by requiring systems to classify English syllogisms as valid or invalid while minimizing content effect (CE), measured as the systematic accuracy gap across plausibility conditions. Systems are ranked by a combined metric that jointly rewards high accuracy and low content bias

[1] All ablation and multi-seed experiments are conducted on the validation split; the official test set is used only for the submitted system.

[2] The premises establish that every city is a location and every location is a capital city, which entails that every city is a capital city - not the reverse.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2512–2519

(see Section 2.1 for details).

We present MLA-CI (Multi-Layer Adversarial for Content Invariance), a DeBERTa-v3-base (He et al., 2021, 2023) classifier with five components (Figure 1): multi-layer feature extraction from early, middle, and late transformer layers; gradient-reversal adversarial training (Ganin and Lempitsky, 2015) against a plausibility discriminator; structure-preserving template augmentation via entity swapping; $2\times$ oversampling of implausible examples; and focal loss (Lin et al., 2017). Our submitted MLA-CI system performed competitively on the official test set; full results and per-condition breakdowns are reported in Section 5.1.

Our principal contribution is a systematic ablation study revealing that *adversarial training is counterproductive* for this task. Across three random seeds, disabling gradient reversal improves the mean validation score from 26.41 ± 0.99 to 38.15 ± 5.32, with accuracy rising from 82.4% to 91.2%. Per-condition analysis reveals that gradient reversal over-suppresses plausibility-correlated features, causing the model to reject valid syllogisms that happen to sound plausible—an *inverted* content effect relative to the classic human belief bias illustrated in Table 1. Meanwhile, template augmentation and multi-layer features are each essential, and removing either causes severe degradation. These findings suggest that data-level debiasing through structure-preserving augmentation is both more effective and more robust than model-level adversarial debiasing for discriminative syllogistic reasoning.

2 Background

2.1 Task Description

SemEval-2026 Task 11 Subtask 1 asks systems to classify syllogisms as valid or invalid. Each example consists of two premises and a conclusion presented in English natural language. The training set contains 960 syllogisms balanced across validity (480 valid, 480 invalid) and plausibility (474 plausible, 486 implausible), yielding four conditions: plausible-valid (PV), plausible-invalid (PI), implausible-valid (IV), and implausible-invalid (II). The test set contains 191 examples. The ranking metric is:

$$\text{Score} = \frac{\text{Accuracy}}{1 + \log(1 + \text{CE})} \tag{1}$$

where CE captures the average accuracy gap attributable to plausibility, computed from both intra-plausibility and cross-plausibility accuracy differences across the four conditions.

2.2 Related Work

Content effects in syllogistic reasoning have been documented in both humans and LLMs. Dasgupta et al. (2022) showed that LLMs exhibit human-like belief biases, and Eisape et al. (2024) provided a systematic comparison across architectures. Bertolazzi et al. (2024) found that supervised fine-tuning can partially mitigate these biases, while Ozeki et al. (2024) and Wysocka et al. (2025) developed evaluation benchmarks for syllogistic reasoning biases.

On mitigation, Valentino et al. (2025) proposed activation steering on Llama-3, Kim et al. (2025) discovered mechanistic circuits for syllogistic inference, and Maraia et al. (2026) introduced Abstract Activation Spaces for content-invariant reasoning all operating on large generative models via model-level interventions. In the NLI debiasing literature, Zhou and Bansal (2020) compared data-level and model-level debiasing for lexical biases. Our work differs by comparing these paradigms specifically for syllogistic content effects in a discriminative classifier, finding that structure-preserving augmentation outperforms adversarial training.

3 System Overview

MLA-CI consists of five components built on a DeBERTa-v3-base encoder (Figure 1). We describe each component and its motivation below.

3.1 Multi-Layer Feature Extraction

Rather than using only the final `[CLS]` representation, we concatenate the `[CLS]` token from three layers of DeBERTa: layer 2 (early), layer 6 (middle), and layer -2 (penultimate). This is motivated by probing studies showing that different transformer layers encode different types of linguistic information (Jawahar et al., 2019; Tenney et al., 2019): early layers capture surface features, middle layers encode syntactic structure, and late layers specialize in task-specific semantics. For syllogistic reasoning, where the model must distinguish logical structure from surface plausibility, accessing multiple levels of representation provides richer features. The concatenated representation has dimensionality $3 \times 768 = 2304$.

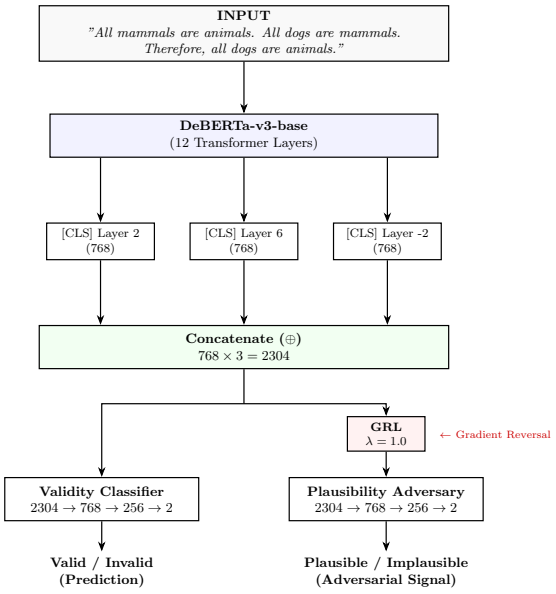

Figure 1: MLA-CI architecture. [CLS] tokens from layers 2, 6, and −2 are concatenated and fed to a validity classifier and an adversarial plausibility discriminator via a gradient reversal layer (GRL).

3.2 Adversarial Plausibility Training

To encourage the shared representation to be invariant to content plausibility, we employ a gradient reversal layer (Ganin and Lempitsky, 2015) connecting the feature representation to a plausibility adversary. During the forward pass, the adversary predicts whether the syllogism is plausible or implausible. During backpropagation, the gradient reversal layer negates the adversary's gradients (scaled by $\lambda = 1.0$), pushing the encoder to produce features from which plausibility cannot be predicted. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{validity}} + \lambda \cdot \mathcal{L}_{\text{adversary}} \quad (2)$$

where the adversary's gradients are reversed before reaching the encoder.

3.3 Template Augmentation

We perform structure-preserving data augmentation by identifying entities in each syllogism and replacing them with entities from the same semantic category (e.g., replacing "dog" with "rabbit" within the *animals* category). This preserves the logical structure—the quantifiers, premise–conclusion relationships, and validity—while varying the content. Entity pools span seven categories (vehicles, buildings, animals, people, objects, food, nature), and we generate up to 2 variations per training example, expanding the training set from 816 to 2,040 samples (2.50×). Critically, augmentation is applied *only* to the training split; the validation set remains unaugmented to ensure fair evaluation.

3.4 Oversampling and Focal Loss

To address the slight plausibility imbalance, we apply 2× duplication of implausible training examples after augmentation, yielding 3,045 training samples. We use focal loss (Lin et al., 2017) with $\gamma = 2.0$ for validity classification, which down-weights well-classified examples to focus on hard cases: $\mathcal{L}_{\text{focal}} = -(1 - p_t)^{\gamma} \log(p_t)$.

4 Experimental Setup

4.1 Data Splits

We split the original 960 training examples into 816 for training and 144 for validation (85/15 split) using stratified sampling across the four plausibility×validity conditions, with a fixed random seed of 42. Template augmentation and oversampling are applied exclusively to the training portion. The 191-example official test set is used only for the final submitted predictions.

4.2 Training Details

We fine-tune DeBERTa-v3-base[3] (184M parameters; 188M with MLA-CI classifier heads) using AdamW (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5}, weight decay 0.01, and linear warmup over 10% of training steps. We train for 4 epochs with batch size 4 and gradient accumulation over 4 steps (effective batch size 16). Gradient clipping is set to 1.0. All experiments use a single NVIDIA Tesla T4 GPU (16GB). Each training run takes approximately 28 minutes for the full pipeline (3,045 samples) and 8 minutes for the vanilla baseline (816 samples). Full hyperparameters are listed in Appendix A.

4.3 Evaluation

We report accuracy, content effect (CE), and the official combined score (Equation 1). We additionally report per-condition accuracy across the four plausibility×validity conditions (PV, PI, IV, II) to diagnose content-dependent behavior. For the key comparison between the full system and the best ablation, we report results across three

[3] https://huggingface.co/microsoft/deberta-v3-base

Configuration	Acc	CE	Score
Vanilla DeBERTa	81.9	5.7	28.17
Full MLA-CI (submitted)	83.3	8.3	25.77
− Adversarial (λ=0)	**93.1**	**4.2**	**35.12**
− Focal loss	90.3	4.4	33.51
− Oversampling	86.1	5.6	29.90
Aug + MultiLayer only	86.8	5.8	29.71
− Augmentation	72.9	14.4	19.51
− Multi-layer	71.5	20.1	17.67

Table 2: Ablation results on held-out validation (144 samples, seed 42). The top block shows baselines; the middle block shows configurations that improve over the full system; the bottom block shows essential components whose removal causes degradation. Bold indicates the best configuration. We denote the removal of a component using the minus (−) prefix.

random seeds (42, 43, 44) as mean ± standard deviation. All ablation and multi-seed experiments are conducted on the held-out validation set; the official test set is used only for the submitted system.

5 Results

5.1 Official Test Results

Our submitted MLA-CI system (all five components active) achieved 79.06% accuracy and 4.17% content effect on the official test set, yielding a combined score of 29.92 (35th of 45 teams).

5.2 Ablation Study

To understand the contribution of each component, we conduct a leave-one-out ablation on the validation set (Table 2). Each row removes a single component while keeping the others intact. We include a vanilla DeBERTa-v3-base baseline (fine-tuned with standard cross-entropy on the original 816 training samples, single `[CLS]` layer, no augmentation or other additions) as a lower reference point.

Four findings emerge from the ablation:

Multi-layer features and augmentation are essential. Removing either causes the largest degradation, dropping the score well below the vanilla baseline. Without multi-layer extraction, accuracy drops to 71.5% and CE rises to 20.1%, indicating that the single final-layer `[CLS]` representation conflates content and structure. Without augmentation, accuracy drops to 72.9% with CE of 14.4%, confirming that structure-preserving data diversity is critical for generalization.

	− Adversarial			Full MLA-CI		
Seed	**Acc**	**CE**	**Score**	**Acc**	**CE**	**Score**
42	93.1	4.2	35.12	83.3	8.3	25.77
43	90.3	4.4	33.69	78.5	5.2	27.81
44	90.3	1.7	45.63	85.4	9.3	25.66
Mean	91.2	3.4	38.15	82.4	7.6	26.41
±Std	±1.3	±1.2	±5.32	±2.9	±1.8	±0.99

Table 3: Multi-seed comparison on validation. The −Adversarial configuration outperforms the Full system on every seed. The lowest −Adversarial score (33.69) exceeds the highest Full system score (27.81) by 5.9 points. We denote the removal of a component using the minus (−) prefix.

Adversarial training is counterproductive. Removing the gradient reversal adversary produces the single largest improvement: accuracy rises from 83.3% to 93.1%, CE decreases from 8.3% to 4.2%, and the combined score increases by +9.35 points. This is the opposite of the intended effect.

The full system underperforms a vanilla baseline. The vanilla DeBERTa baseline (28.17) outperforms the full five-component system (25.77), indicating that the adversarial component is so harmful that it more than cancels the gains from the other four components.

Focal loss and oversampling interact with adversarial training. Removing adversarial training alone (keeping focal loss and oversampling) yields 35.12, whereas removing all three yields only 29.71. This indicates that focal loss and oversampling contribute positively in the absence of adversarial training, but their benefits are masked or inverted when gradient reversal is active.

5.3 Multi-Seed Robustness

To verify that the adversarial training finding is not an artifact of a single random seed, we run both the full system and the −Adversarial configuration across three seeds (42, 43, 44). Table 3 reports the results.

The finding is robust: removing adversarial training improves the score on *every* seed, with a mean improvement of +11.74 points. The gap is large enough that the worst −Adversarial run (33.69, seed 43) exceeds the best full-system run (27.81, seed 43) by 5.9 points, confirming that the effect is not attributable to random variation. Notably, the full system exhibits higher accuracy

Config	PV	PI	IV	II	σ
Vanilla	75.0	85.7	80.6	86.5	4.61
Full system	75.0	80.0	91.7	86.5	6.31
− Adversarial	**88.9**	**91.4**	**94.4**	**97.3**	**3.16**
− Augment.	88.9	60.0	80.6	62.2	12.21
− Multi-layer	61.1	82.9	80.6	62.2	10.08

Table 4: Per-condition accuracy (%) on validation (seed 42) for key configurations. PV = Plausible-Valid, PI = Plausible-Invalid, IV = Implausible-Valid, II = Implausible-Invalid. σ = standard deviation (lower is more uniform).

variance ($\pm$2.9 vs. $\pm$1.3), suggesting that adversarial training also introduces training instability.

5.4 Per-Condition Error Analysis

Table 4 breaks down accuracy by the four plausibility$\times$validity conditions for key configurations (seed 42).

The full system's weakest condition is *Plausible-Valid* (75.0%), meaning it frequently rejects valid syllogisms that happen to sound plausible. This is the *inverse* of the classic human belief bias, which tends to accept plausible conclusions regardless of validity (Dasgupta et al., 2022). Removing adversarial training raises PV accuracy by 13.9 percentage points (from 75.0% to 88.9%), the largest single-condition improvement across all ablations, and produces the most uniform accuracy profile ($\sigma = 3.16$).

This pattern has a clear explanation: the gradient reversal layer, by explicitly penalizing features correlated with plausibility, causes the encoder to suppress exactly those features that happen to co-occur with valid logical structure in the plausible-valid condition. The model learns to be *skeptical* of plausible-sounding arguments, over-correcting beyond content invariance into active content aversion.

5.5 Adversarial Strength Sensitivity

To verify that the adversarial training finding is not an artifact of the single value $\lambda = 1.0$ used in the submitted system, we sweep $\lambda \in \{0.0, 0.05, 0.1, 0.3, 0.5, 1.0, 2.0\}$ with all other components fixed (seed 42). Table 5 reports the results.

Three patterns emerge from the sweep. First, accuracy peaks in the low-λ range: $\lambda = 0.1$ achieves the highest accuracy (93.75%) and $\lambda = 0.05$ achieves the highest accuracy-based score when CE is also considered (34.82), both exceeding the $\lambda = 0.0$ baseline (31.58). This suggests that very light adversarial pressure can provide a marginal regularization benefit. Second, the relationship between λ and accuracy is non-monotonic: performance degrades sharply beyond $\lambda = 0.3$ (accuracy drops from 93.06% to 84.72% at $\lambda = 0.5$), confirming that standard-strength adversarial training is harmful. Third, $\lambda = 2.0$ causes near-complete model collapse (53.47% accuracy, near-chance), demonstrating that strong gradient reversal catastrophically disrupts learning.

λ	Acc	CE	Score	Best Ep
0.0	90.97	5.56	31.58	4
0.05	**93.06**	4.33	34.82	2
0.1	**93.75**	5.71	32.28	4
0.3	93.06	7.14	30.05	3
0.5	84.72	5.56	29.41	1
1.0	75.69	**1.75**	**37.66**	2
2.0	53.47	23.95	12.68	3

Table 5: λ sweep results on validation (seed 42). Accuracy peaks at low λ values (0.05–0.1) but degrades sharply beyond $\lambda = 0.3$. At $\lambda = 1.0$, accuracy drops to 75.69% though CE reaches its minimum (1.75%) because the model predicts near-uniformly. At $\lambda = 2.0$, the model collapses to near-chance.

Notably, $\lambda = 1.0$ achieves the highest combined score (37.66) due to an extremely low CE of 1.75%, but this reflects near-uniform predictions across conditions (all four condition accuracies within 75.0–77.8%) rather than genuinely content-invariant reasoning—the model has sacrificed discriminative ability for apparent fairness.

These results refine our original finding: the adversarial component is not universally harmful, but the submitted value ($\lambda = 1.0$) falls well into the over-correction regime. Optimal performance requires $\lambda \leq 0.1$, with diminishing and then negative returns beyond that threshold.

6 Discussion

Our results highlight a tension between two debiasing paradigms. Data-level debiasing via augmentation exposes the model to diverse surface realizations of the same logical structures, allowing it to learn structure-sensitive features organically. Model-level debiasing via adversarial training explicitly penalizes plausibility-predictive features, but cannot distinguish features incidentally correlated with plausibility from those causally necessary for validity judgments. In the plausible-valid

condition, features indicating plausibility may overlap with features signaling validity (e.g., coherent term relationships), and suppressing them degrades performance. The λ sweep (Table 5) confirms this: only very light adversarial pressure ($\lambda \leq 0.1$) preserves accuracy, while standard or strong values ($\lambda \geq 0.5$) cause substantial degradation.

This complements work by Valentino et al. (2025) and Maraia et al. (2026), whose model-level interventions target generative LLMs with billions of parameters. Our results suggest that for discriminative classifiers with limited training data (∼1,000 examples), adversarial objectives risk over-correction, and data-level augmentation provides a simpler, more effective alternative.

Classification-based approach. Our system treats syllogistic reasoning as a text classification task, concatenating premises and conclusion without explicit modeling of logical structure (e.g., syllogistic figures or premise–conclusion parsing). This is a deliberate design choice to isolate the effect of debiasing strategies in a controlled setting. However, it limits the system to Subtask 1 and may miss structural cues that explicit logical modeling could capture. Future work could incorporate structure-aware encoding, such as separate premise and conclusion representations with cross-attention, to jointly address validity classification and premise extraction.

Limitations. Our validation set contains 144 examples (∼36 per condition), so per-condition estimates remain noisy despite multi-seed robustness. The post-hoc finding has not been verified on the official test set. While we explored seven λ values (Section 5.5), we did not test annealing schedules, variation in adversary capacity (e.g., deeper or wider discriminator architectures), or curriculum/staged training strategies, any of which might recover adversarial training's effectiveness.

7 Conclusion

We presented MLA-CI for SemEval-2026 Task 11 Subtask 1, combining multi-layer feature extraction, adversarial training, template augmentation, oversampling, and focal loss. Through ablation confirmed across three seeds, we found that adversarial gradient reversal at the submitted strength ($\lambda = 1.0$) is actively harmful—over-suppressing content-correlated features and reducing plausible-valid accuracy. Removing it yields a mean score of 38.15 ± 5.32 vs. 26.41 ± 0.99 for the full system. A sweep over seven λ values reveals that only very light adversarial pressure ($\lambda \leq 0.1$) preserves accuracy, with performance degrading sharply beyond $\lambda = 0.3$ and collapsing at $\lambda = 2.0$. When structure-preserving augmentation provides sufficient content diversity, standard-strength adversarial debiasing degrades performance through over-correction. Future work could explore λ annealing schedules, adversary capacity variation, structure-aware encoding with explicit logical parsing, LLM-generated augmentation, and cross-task generalization to other reasoning benchmarks.

References

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd van 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 (NAACL)*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics.

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.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

G. 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*.

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*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the 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)*.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and André 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 (NAACL)*.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Appendix

A Hyperparameters

Table 6 summarizes all hyperparameters used in training.

Hyperparameter	Value
Base model	DeBERTa-v3-base
Max sequence length	512
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	4
Gradient accum. steps	4
Effective batch size	16
Epochs	4
Warmup ratio	10%
Gradient clip	1.0
Dropout	0.2
λ_{adv}	$\{0.0, 0.05, 0.1, 0.3, 0.5, 1.0, 2.0\}$ (sweep)
Focal loss γ	2.0
Oversampling ratio	$2\times$ (implausible)
Augmentation variants	2 per sample
Random seeds	42, 43, 44

Table 6: Training hyperparameters.

B Multi-Layer Feature Extraction Details

The three DeBERTa layers used for feature extraction are:

- **Layer 2** (early): captures lexical and surface-level features
- **Layer 6** (middle): encodes syntactic structure and compositional semantics
- **Layer −2** (penultimate): provides task-relevant abstract representations while avoiding over-specialization of the final layer

C Epoch-Level Details

Table 7 shows the best-epoch selection for each ablation configuration (seed 42). Most configurations reach their best combined score at epoch 2 or 4, confirming that 4 epochs is sufficient for convergence.

D Augmentation Details

The template augmenter identifies entities from seven semantic categories and replaces them with randomly sampled alternatives from the same category, preserving logical structure. Table 8 shows the entity pools.

Configuration	Best Epoch	Train Acc (%)
Vanilla DeBERTa	4	90.93
Full MLA-CI	4	98.92
− Adversarial	2	97.08
− Focal loss	4	99.47
− Oversampling	4	95.83
Aug + MultiLayer only	4	98.19
− Augmentation	4	89.99
− Multi-layer	3	73.60

Table 7: Best epoch and corresponding training accuracy for each ablation configuration (seed 42).

Category	Entities
Vehicles	car, bicycle, motorcycle, truck, bus, train, airplane, boat
Buildings	house, building, tower, castle, barn, shed, cabin, mansion
Animals	dog, cat, horse, bird, fish, rabbit, lion, tiger
People	person, human, individual, citizen, student, teacher, worker, doctor
Objects	book, table, chair, tool, device, machine, instrument, gadget
Food	fruit, vegetable, meat, drink, meal, snack, dish, dessert
Nature	tree, plant, flower, rock, mountain, river, ocean, forest

Table 8: Entity pools used for template augmentation. Replacements are drawn from the same category as the original entity.

E Reproducibility and Computational Environment

All experiments were conducted on Google Colab using a single NVIDIA Tesla T4 GPU (16GB) with CUDA, PyTorch, and Transformers 4.47.1. We use fixed random seeds (42, 43, 44) for Python, NumPy, and PyTorch; however, we do not enforce full GPU determinism (i.e., `torch.backends.cudnn.deterministic` is not set), as this substantially increases training time. Consequently, results may vary by 1–2 percentage points across re-runs in different computational environments due to non-deterministic GPU operations, CUDA version differences, and floating-point accumulation order. The official test score (29.92) was obtained during the evaluation phase; subsequent re-runs of the same configuration on the validation split produced scores within ± 1 point of the reported values (Table 3). Our multi-seed analysis (seeds 42, 43, 44) is designed to quantify this variance: the key finding that removing adversarial training improves performance holds consistently across all seeds, with non-overlapping score ranges.