HABIB_TAZ at SemEval-2026 Task 11: Disentangling Formal Logic from Content via Synthetic Training and Multi-Objective Optimization

Abdullah Shaikh*, **Zain Naqi***, **Taha Zahid***, **Sandesh Kumar***, **Abdul Samad**
Dhanani School of Science & Engineering
Habib University, Pakistan
{zn09224, as09245, tz09220}@st.habib.edu.pk
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk

Abstract

While Large Language Models (LLMs) excel in many general NLP tasks, their formal reasoning capabilities are often compromised by content effects, demonstrating a measurable bias towards real-world plausibility. In this paper, we present our system for SemEval-2026 Task 11, which evaluates the ability of models to disentangle formal logic from content across 12 languages with and without distractor premises. We address this challenge using mDeBERTa-v3 networks fine-tuned on a synthetic, rule-based dataset of syllogistic schemes to avoid the semantic noise of LLM-augmented data. To explicitly decouple plausibility from logical structure, our training pipeline employs a multi-objective loss function combining Adaptive Group Distributionally Robust Optimization (DRO), a scheduled differentiable bias penalty, and KL-Divergence consistency regularization. Our system achieved #1 ranks and perfect Ranking Scores (100.0) with 0.00% bias and 100.0% accuracy on Subtask 1 (English), Subtask 2 (Noisy English), and Subtask 3 (Multilingual). On the highly complex Subtask 4 (Noisy Multilingual), the system achieved the 6th rank with 89.06% Accuracy and F1-score, alongside a limited 2.89% Bias and a 37.78 Ranking Score. Our dataset generation engine and codebase are publicly available to facilitate future work on robust logical reasoning.

1 Introduction

While Large Language Models (LLMs) (Cheng et al., 2025; Wan et al., 2025; Quan et al., 2024) excel in general NLP, they struggle with "content effects", overestimating the logical validity of arguments aligned with world knowledge (Dasgupta et al., 2024; Valentino et al., 2025; Kim et al., 2024). This entanglement hinders the development of trustworthy AI that prioritizes logical robustness (Huang et al., 2024). We address this in SemEval-2026 Task 11 (Valentino et al., 2026), aiming to assess the formal validity of syllogistic arguments across 12 diverse languages[1], independent of their plausibility (alignment with world knowledge). Furthermore, the task requires models to exhibit robustness by identifying relevant premises amidst irrelevant noise.

Our system utilizes mDeBERTa-v3 backbones (He et al., 2021) pre-trained on NLI tasks (Laurer et al., 2022). To ground the architecture in formal logic, we curate a synthetic dataset via a rule-based engine following syllogistic schemes (Bertolazzi et al., 2024), avoiding the noise of LLM-augmented data. We optimize via a multi-objective function integrating Group DRO (Sagawa et al., 2020), a scheduled bias-weighting parameter λ, and KL Divergence to decouple formal reasoning from plausibility, and to enforce structural invariance across complex linguistic variations.

Our code and data generation scripts are publicly available to facilitate further research.[2]

2 Related Work

Recent work on logical reasoning in LLMs can be categorized along four directions, following the taxonomy of (Cheng et al., 2025).

Neuro-Symbolic Approaches (Quan et al., 2024) integrates external theorem provers to verify LLM-generated natural language explanations. By iteratively formalizing explanations and validating them with a logic solver, their feedback loop ensures logical soundness and completeness.

Prompt-Based Methods (Wan et al., 2025) propose a Syllogistic-Reasoning Framework of Thought (SR-FoT), extending Chain-of-Thought

* Equal contribution.

[1] English (en), German (de), Spanish (es), French (fr), Italian (it), Dutch (nl), Portuguese (pt), Russian (ru), Chinese (zh), Swahili (sw), Bengali (bn), and Telugu (te).

[2] https://github.com/TahaZahid05/SemEval26-Task-11

Quadrant	Example (Premises → Conclusion)	Validity	Plausibility
Valid Plausible	All cats are animals. Luna is a cat → **Luna is an animal.**	1	1
Valid Implausible	All birds can swim. Eagles are birds → **Eagles can swim.**	1	0
Invalid Plausible	All dogs have fur. Some pets have fur → **Some pets are dogs.**	0	1
Invalid Implausible	All rocks are soft. All clouds are soft → **Rocks are clouds.**	0	0

Table 1: Task structure and logical quadrants. In "Noisy" tracks (ST2/ST4), models must also identify the subset of relevant premises, filtering out distractor sentences like "The sky is blue."

prompting to a multi-stage syllogistic reasoning process. By generating major and minor premises before the conclusion, SR-FoT enforces strict information control, reducing errors and improving logical consistency.

Model Optimization (Valentino et al., 2025) proposes Activation Steering to disentangle content plausibility from logical validity within LLM hidden representations. Using probing techniques to identify content-sensitive layers, they apply conditional steering to guide the model toward formal reasoning. Similarly, (Wu et al., 2024) proposes a Logical Control Framework that separates content and logic subspaces via contrastive learning, nudging representations toward logical validity during inference without altering content.

Analytical Studies (Kim et al., 2024) identifies distinct attention circuits responsible for formal reasoning versus content-dependent processing, revealing that certain attention heads are both necessary and sufficient for logical inference. On the evaluation side, (Patel et al., 2024) introduces Multi-LogiEval, demonstrating that LLM accuracy degrades as the number of premises and reasoning steps increases.

3 Background and Task Description

The task requires classifying the logical validity of a conclusion (C) given premises (P) across four subtasks. Evaluation targets the 2×2 intersection of **Validity** and **Plausibility** (Table 1).

3.1 Subtasks and Metrics

We participated in all four subtasks: **ST1** (English), **ST2** (English with irrelevant premises), **ST3** (12-language multilingual), and **ST4** (12-language multilingual with irrelevant premises). ST2 and ST4 require joint validity classification and relevant premise selection. The primary metric is the **Ranking Score** (S_{rank}), which divides a performance component P by a logarithmic penalty on the **Total Content Effect (TCE)** (Valentino et al., 2026). P is Accuracy for ST1/ST3 and the average of Accuracy and F1 for ST2/ST4. TCE measures bias by comparing accuracy across the validity–plausibility quadrants; lower TCE indicates higher logical robustness.

3.2 Dataset

The training dataset provided contained 960 English data points balanced across quadrants (VP: 25.0%, VI: 25.0%, IP: 24.4%, II: 25.6%).

4 System Overview

Our system is designed to decouple semantic patterns from logical structure through synthetic data generation, bias-aware optimization, and leveraging the structural attention of mDeBERTa-v3 backbones.

4.1 Architecture

We utilize mDeBERTa-v3 (He et al., 2021) as our backbone, initialized from NLI-pretrained checkpoints (Laurer et al., 2022): `mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` for ST1 and ST3, `DeBERTa-v3-large-mnli-fever-anli-ling-wanli` for ST2, and `microsoft/mdeberta-v3-base` for ST4.[3] The larger backbone for ST2 proved necessary for disentangling noisy premises. For ST1/ST3, a classification head consumes the `[CLS]` token for binary validity classification. For ST2/ST4, we use two specialized heads:

[3]All checkpoints are available on HuggingFace under the `MoritzLaurer` namespace, except for ST4 which uses a Microsoft checkpoint.

- **Classification Head:** Predicts logical validity from the [CLS] embedding.
- **Premise Selection Head:** Uses span-based pooling between [SEP] tokens to extract premise representations. We apply Masked Mean Pooling to each span, passing the resulting fixed-size vectors to a sigmoid layer for relevance probability.

4.2 Data Strategy

To overcome the scarcity of the provided dataset ($N = 960$), we developed a rule-based engine to generate synthetic data based on the 256 Aristotelian moods and figures (Bertolazzi et al., 2024).

Vocabulary and Plausibility We select nouns from WordNet (Miller, 1995) hypernym/hyponym chains using the NLTK library (Bird et al., 2009). Plausible samples preserve hierarchical ordering (e.g., $car \subset vehicle \subset artifact$) for nouns, while implausible samples use cross-chain or random nouns. To further reduce content reliance, we introduce **Symbolic Data** (using gibberish/symbolic variables) and **Complex Rephrasing** (e.g., mapping "All x are y" to "Every x is, without exception, a y.").

Relevant Premises Labeling For invalid syllogisms in ST2 and ST4, we retain the indices of premises that provide partial entailment. This prevents the model from treating premise relevance as a purely binary consequence of overall validity, encouraging structural over-fitting to logical forms.

Multilingual Expansion We first generated English samples, then translated nouns via DeepL[4], and then translated the structure using language-specific rule-based templates. This ensures that the quantifiers (e.g., "No", "Some") are not modified across all 12 target languages.

Syntactic Robustness (HANS) To ensure the models do not exploit surface-level syntactic heuristics (e.g., lexical overlap) during inference, we additionally integrate samples from the Heuristic Analysis for NLI Systems (HANS) dataset (McCoy et al., 2019) into our robust training distributions.

4.3 Bias-Aware Optimization

To optimize for the Ranking Score, we use a multi-objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{DRO} + \lambda_{bias}\mathcal{L}_{bias} + \gamma\mathcal{L}_{KL} \quad (1)$$

[4] https://www.deepl.com/

Adaptive Group DRO and Scheduling Instead of Cross Entropy Loss, which prioritizes global performance, we utilize **Group Distributionally Robust Optimization (DRO)** (Sagawa et al., 2020) to ensure that the model stays consistent by focusing more on low-performing subgroups. We define $G = 6$ subgroups: $\{VP, VI, IP, II\}$ and two *Symbolic* groups. For multilingual tasks (ST3/ST4), each of the six subgroups is further divided into 12 subgroups, one for each language, resulting in a total of 72 subgroups. The model optimizes $\mathcal{L}_{DRO} = \sum_{g=1}^{G} q_g \mathcal{L}_g$, where group weights q_g are updated adaptively:

$$q_g^{(t+1)} \propto q_g^{(t)} \exp(\eta \mathcal{L}_g^{(t)}) \quad (2)$$

To ensure the model establishes a logical baseline before unbiasing, we implement **Dynamic Scheduling**: λ_{bias} and the DRO step-size η are initialized at 0 for the first epoch and incrementally increased. This prevents the optimization from collapsing into sub-optimal local minima where the model might ignore semantic content before mastering formal syllogistic structures.

Differentiable Bias Penalty To align the training objective with the Content Effect (CE) evaluation metrics, we formulate a loss component using sigmoid confidence scores $\hat{y}$. Let $\hat{y}_{v,pl}$ be the model's confidence scores for samples with validity $v \in \{0, 1\}$ and plausibility $pl \in \{0, 1\}$. We calculate:

$$\bar{\hat{y}}_{Intra} = \frac{1}{2} \sum_{p \in \{0,1\}} |\text{mean}(\hat{y}_{1,p}) - \text{mean}(\hat{y}_{0,p})|$$

$$\bar{\hat{y}}_{Cross} = \frac{1}{2} \sum_{v \in \{0,1\}} |\text{mean}(\hat{y}_{v,1}) - \text{mean}(\hat{y}_{v,0})|$$

The final $\mathcal{L}_{bias}$ is the mean of these effects. This approach allows the gradient to penalize uneven confidence distributions across quadrants, effectively "forcing" the model to maintain high logical certainty regardless of real-world plausibility.

Consistency via KL-Divergence To ensure robustness to linguistic variation, we minimize the KL-Divergence ($\mathcal{L}_{KL}$) (Kullback and Leibler, 1951) between the output distributions of a simple syllogism (teacher) and its complex rephrasing (student). For example, mapping the simple form "All A are B" to the complex "If something is A, then it is B". This $\min \text{KL}(P_{simple} || P_{complex})$ optimization forces the model to align the semantic

representation of complex conditionals with their fundamental logical forms, thereby preventing reliance on lexical shortcuts.

4.4 Data Splits and Preprocessing

We use a stratified 80/10/10 split (Train/Validation/Test) for both the training dataset provided by the curators (Valentino et al., 2026), and our augmented dataset. Stratification is performed across to ensure balanced logical and semantic representation in every split.

4.5 Dataset Composition

Our augmentation pipeline expanded the fully balanced initial dataset ($N = 960$) to include a more complete set of logical forms and variations: $23,524$ total samples for ST1, $27,988$ for ST2, 388,992 for ST3, and $648,632$ for ST4. All subsets remain balanced across the four logic/plausibility quadrants and also across languages for multilingual subtasks.

4.6 Hyperparameters

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear schedule and a warmup ratio of 0.06. To account for the stability of the backbone, we apply a differential learning rate ($\text{LR}_{backbone} = 0.1 \times \text{LR}_{head}$). Specific subtask hyperparameter configurations are detailed in Appendix C.

4.7 Implementation and Hardware

The system is implemented in PyTorch (Paszke et al., 2019) using HuggingFace Transformers (Wolf et al., 2020). Training used NVIDIA Tesla P100 (16GB) and T4 Tensor Core (16GB) GPUs with Mixed Precision (FP16) and a Gradient Scaler.

5 Results and Analysis

5.1 Main Quantitative Findings

The performance of our best models is summarized in Table 2, with the complete ablation variations available in Appendix D (Table 7). Our primary objective was to maximize the Ranking Score (S_{rank}) by minimizing the Total Content Effect (TCE) and maximizing the performance component P. We additionally explored a couple of other experimental loss components, including **Mean Squared Error (MSE)** and **FreeLB** (Zhu et al., 2019); while these provided competitive stability, they were not included in the final system overview as they were not in our best results. **Details on these experimental components can be found in Appendix A.**

As shown in the ST1 results, all configurations achieved perfect performance, suggesting that for ST1, the mDeBERTa-v3 backbone combined with our augmented dataset is sufficient.

However, the complexity of ST2 revealed performance gaps. Our final configuration using **KL + DRO + Bias** on the large model backbone achieved a perfect score of 100.0, successfully disentangling formal reasoning from noisy premises and content bias.

For ST3, using the bias alone produced near-perfect accuracy. Both **Bias + HANS** and **Bias + HANS + DRO** achieved a perfect score of 100.00. The addition of KL or MSE regularization failed to preserve this performance, suggesting that bias loss combined with HANS (with or without DRO) is sufficient to disentangle reasoning from content bias.

For the highly complex ST4, the **Bias** configuration achieved a Ranking Score of 37.78 during official evaluation. Notably, post-competition analysis revealed a drastic performance gap when evaluating on the original multilingual test set compared to its English translation generated via ChatGPT (OpenAI, 2026). As shown in Table 7 (Appendix D), the **DRO + Bias** model scored 25.50 on multilingual data but jumped to 40.35 on translated English data. A similar $\approx 60\%$ rank increase occurs for the standalone **Bias** model. Interestingly, as we discuss further in Section 5.2, applying KL-Divergence consistency regularization significantly closes this cross-lingual gap.

5.2 Ablation and Error Analysis

For ST2, our experiments indicate that while the Vanilla model was able to achieve some success, it is susceptible to content effects (2.13% TCE).

Surprisingly, adding only Differentiable Bias (40.13) or DRO + Differential Bias (36.42) degraded the performance compared to vanilla. DRO + Differential Bias showed the highest content effect (4.26%), which suggests that aggressive optimization can dominate the loss function and degrade performance. This finding underscores the importance of carefully designed multi-component systems rather than isolated regularization techniques. We found that an additional component, such as KL-Divergence or MSE, acted as a regularization anchor ensuring that even as the model is

Track	Best Configuration	Acc	F1	TCE	S_{rank}
ST1	KL + DRO + Bias	100.0	–	0.0	100.0
ST2	KL + DRO + Bias (Large)	100.0	100.0	0.0	100.0
ST3	Bias + HANS	100.0	–	0.0	100.0
ST4	Bias (ET)	89.06	90.62	2.21	41.43

Table 2: Best system performance per subtask. Full ablation configurations are available in Appendix D (Table 7).

pushed to be unbiased, it remains grounded in the underlying formal syllogistic structure.

Our complete system (KL + DRO + Bias) achieved a 57.53 ranking score on the base model, representing a 25% improvement over vanilla. The FreeLB + MSE + DRO + Bias configuration achieved identical performance, indicating that both KL divergence and FreeLB provide similar consistency regularization benefits. However, FreeLB is computationally more expensive since it runs multiple forward passes per input.

Scaling the architecture from the Base to the Large model in the ST2 (KL + DRO + Bias) configuration yielded a perfect ranking score. This suggests that the capacity to abstract logical relationships from noisy premises scales directly with model parameters. Furthermore, while MSE and FreeLB provided stability during baseline experimentation, KL-Divergence consistency regularization proved more efficient and equally effective when generalizing to the Large backbone.

For ST3, the Vanilla model achieved a strong accuracy of 91.15%, but remained highly susceptible to content effect (8.29% TCE).

Introducing only the Differentiable Bias significantly improved the performance, accuracy increased to 98.96, while content effect dropped sharply to 1.04%, which suggests the effectiveness of bias-aware loss in mitigating the content effect. However, the most substantial jump occurred with the inclusion of the HANS dataset. The **Bias + HANS** configuration achieved a perfect 100.00 final score. Integrating DRO into this configuration maintained the perfect performance, indicating the model's stability across different data distributions.

Interestingly, unlike the trends observed in ST2, additional consistency regularizers, KL-Divergence and MSE, failed to sustain the perfect score. Instead, they degraded the ranking scores and reintroduced the content effect. Both KL and MSE yielded identical results, suggesting that the model had already reached the optimal state and additional regularization may have introduced unnecessary constraints.

Furthermore, a critical takeaway from the ST3 results is the comparison between models trained without symbolic data and those trained with symbolic data. The models trained with symbolic data consistently outperformed the ones trained without it, suggesting that symbolic datasets contributed heavily to generalization. By training on abstract symbols rather than just real-world nouns, the model effectively disentangles logical reasoning from content bias.

For ST4, a pattern similar to ST2 emerges: DRO + Bias and Bias alone perform worse than Vanilla on multilingual test data (MT). To diagnose this, we translated the multilingual test set to English (ET) via GPT-5.2 (OpenAI, 2026). The results reveal a significant cross-lingual gap: DRO + Bias scored 25.50 on MT but 40.35 on ET, and Bias alone scored 25.98 vs. 41.43, a $\approx 60\%$ rank increase, indicating that the core failure stems from (i) difficulty generalizing multilingual premise structures and/or (ii) "translationese" noise in mDeBERTa-v3's pre-training data (Artetxe et al., 2020). In contrast, the **KL + DRO + Bias** configuration achieved near-identical scores on both test sets (S_{rank} of 31.16 on MT vs. 32.29 on ET), showing that KL-Divergence consistency regularization makes the model robust to cross-lingual variation by anchoring on simple English syllogisms and minimizing output divergence across all 12 language variants. However, this requires 24 forward passes per data point; due to compute constraints, training data was reduced to $\approx 4\%$, suggesting further gains are achievable with more data.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

L. Bertolazzi, A. Gatt, A. M. Gliozzo, G. Greco, A. Madotto, and M. Valentino. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *arXiv preprint*, arXiv:2406.11341.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

F. Cheng, H. Li, F. Liu, R. van Rooij, K. Zhang, and Z. Lin. 2025. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint*, arXiv:2502.15652.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. Language models show human-like content effects on reasoning tasks. *Preprint*, arXiv:2207.07051.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, and 51 others. 2024. Trustllm: Trustworthiness in large language models. *Preprint*, arXiv:2401.05561.

G. Kim, M. Valentino, D. Dalal, Z. Zhao, and A. Freitas. 2024. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. *arXiv preprint*, arXiv:2408.08590.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Preprint, Open Science Framework*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Preprint*, arXiv:1902.01007.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

OpenAI. 2026. ChatGPT (gpt-5.2 version). https://chat.openai.com. Large language model accessed via web interface.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

N. Patel, M. Kulkarni, M. Parmar, A. Budhiraja, M. Nakamura, N. Varshney, and C. Baral. 2024. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Available: arXiv:2406.17169.

X. Quan, M. Valentino, L. A. Dennis, and A. Freitas. 2024. Verification and refinement of natural language explanations through llm-Symbolic theorem proving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2933–2958. Available: arXiv:2405.01379.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *Preprint*, arXiv:1911.08731.

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 language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

W. Wan, Z. Yang, Y. Chen, C. Luo, R. Wang, K. Cai, N. Kang, L. Lin, and K. Wang. 2025. Sr-fot: A syllogistic-reasoning framework of thought for large language models tackling knowledge-based reasoning tasks. *arXiv preprint*, arXiv:2501.11599.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

X. Wu, Y. Bu, and 1 others. 2024. Content-free logical modification of large language models by disentangling and modifying logic representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. DOI pending.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *CoRR*, abs/1909.11764.

A Experimental Loss Components

As noted in Section 5.1, while our final winning configuration relied on KL-Divergence for consistency regularization and avoided adversarial perturbations to maintain computational efficiency on the Large backbone, our detailed ablation study (Appendix D) reports metrics on two additional experimental loss components: Mean Squared Error (MSE) consistency and FreeLB (Free Large-Batch Adversarial Training).

A.1 Mean Squared Error (MSE) Consistency Regularization

Before adopting KL-Divergence, we experimented with Mean Squared Error (MSE) to enforce structural invariance. Under this configuration, every data point was structured to contain both a `syllogism_simple` and a `syllogism_complex` text variant. For the multilingual subtasks (ST3/ST4), this was expanded further such that a single data point could contain up to 24 language-specific simple and complex variations of the same underlying logical structure. During the forward pass, the model generated logits for both the simple constraint (z_{simple}) and the complex constraint ($z_{complex}$). The MSE consistency penalty was formulated as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(z_{simple}^{(i)} - z_{complex}^{(i)})^2 \quad (3)$$

This loss component successfully anchored the model, ensuring that the confidence scores for semantically complex variations closely mapped to their fundamental, simple logical counterparts. While MSE provided competitive stability (Ranking Score 45.03 in ST2), we ultimately found that minimizing the KL-Divergence between the output probability distributions was more effective at preventing the model from relying on lexical shortcuts than minimizing the absolute difference between the raw logits.

A.2 Free Large-Batch Adversarial Training (FreeLB)

To further test the robustness of the model against grammatical noise and irrelevant premises, we implemented FreeLB (Zhu et al., 2019). FreeLB adds adversarial perturbations to word embeddings during training, forcing the model to make consistent predictions even when the input representations are subjected to small, continuous gradient-based attacks. We applied the standard FreeLB algorithm, executing multiple forward passes per input batch. In each inner step, we perturbed the embeddings by a constrained magnitude ϵ and calculated the loss, updating the model parameters based on the accumulated gradients of the adversarially perturbed inputs. As shown in Table 7, the combination of FreeLB + MSE + DRO + Bias achieved a Ranking Score of 57.53 on the Base model, which was identical to the performance of our KL + DRO + Bias configuration. FreeLB successfully stabilized the aggressively debiased model. However, because FreeLB requires multiple forward passes per iteration, it is computationally expensive. When transitioning our architecture to the Large mDeBERTa-v3 backbone to maximize the Ranking Score, we dropped FreeLB in favor of the computationally lighter KL-Divergence approach, which yielded the same consistency benefits without multiplying the training time.

B Extended Data Generation Procedure

As described in Section 3.2, our synthetic dataset was generated programmatically. To isolate the effects of structural noise, linguistic complexity, and consistency pairing, the data was generated via independent script executions rather than a monolithic pipeline.

B.1 Base Generation

The baseline dataset was generated using a 256-rule Aristotelian logic engine. This initial procedure outputs 'Simple' English syllogisms encompassing non-symbolic nouns, symbolic variables, and both 3-variable and 4-variable logical constraints. All subsequent augmentation scripts act upon this foundational subset.

B.2 Irrelevant Premise Injection

For robustness configurations (e.g., ST2/ST4), modifier scripts were executed over the base English data to inject structural noise:

1. **Premise Removal:** The modifier script discarded a structurally critical premise from a designated subset of valid chains, directly invalidating the logic.

2. **Distractor Generation:** The sequence then generated logically disjoint distractor premises (e.g., inserting a premise regarding *"clouds"* into a syllogism concerning *"dogs"*) and concatenated them to the input text.

B.3 Paired Semantic Augmentation

Consistency optimization (e.g., KL-Divergence or MSE) requires identically paired "simple" and "complex" variants of underlying logical forms. For English datasets, a secondary complex-rephrasing script was applied to the base output, applying synonym replacement and structural permutations while maintaining parity with the corresponding simple input matrices.

B.4 Multilingual Translation

For the 11 non-English language tracks (ST3/ST4), translation followed two independent pathways depending on the required complexity constraint:

1. **Simple Translation:** The complete English input text (with or without injected noise) was translated into the target languages via the DeepL API.

2. **Complex Translation:** To execute complex rephrasing natively across languages, a specialized script isolated and translated only the semantic nouns from the English text. These localized noun-sets were subsequently passed to 11 language-specific generator scripts (e.g., `ComplexSyllogisms_French.ipynb`), which applied complex phrasing mappings independently. Because symbolic datasets lack semantic noun dependencies, they bypassed this translation step entirely.

C Hyperparameter Configurations

C.1 Subtask 1 (ST1) Configurations

Hyperparameter	Value
Model	mDeBERTa-v3-base
Batch Size	32
Gradient Accumulation Steps	1
Epochs	10
Learning Rate	2×10^{-5}
Warmup Ratio	0.06
Weight Decay	0.01
Max Sequence Length	96
Dropout Rate	0.1
Max λ_{bias}	2.0
γ (KL Consistency)	0.5
DRO Warmup	1 Epoch
Early Stopping Patience	2
Early Stopping Min Δ	0.001
Mixed Precision (FP16)	True

Table 3: Hyperparameters for Subtask 1 (ST1).

C.2 Subtask 2 (ST2) Configurations

Hyperparameter	Value
Model	DeBERTa-v3-large
Batch Size	6
Gradient Accumulation Steps	10
Effective Batch Size	60
Epochs	10
Learning Rate	5×10^{-6}
Warmup Ratio	0.06
Weight Decay	0.01
Max Sequence Length	232
Premise Buffer Size	9
Dropout Rate	0.1
Max λ_{bias}	2.0
γ (KL Consistency)	0.5
DRO Warmup	1 Epoch
Early Stopping Patience	2
Early Stopping Min Δ	0.001
Mixed Precision (FP16)	True

Table 4: Hyperparameters for Subtask 2 (ST2).

C.3 Subtask 3 (ST3) Configurations

Hyperparameter	Value
Model	mDeBERTa-v3-base-xnli
Batch Size	144
Gradient Accumulation Steps	1
Effective Batch Size	144
Epochs	3
Learning Rate	2×10^{-5}
Warmup Ratio	0.06
Weight Decay	0.01
Max Sequence Length	96
Dropout Rate	0.1
Max λ_{bias}	2.5
DRO Warmup	1 Epoch
Early Stopping Patience	2
Early Stopping Min Δ	0.001
Mixed Precision (FP16)	True
Pooling type	cls

Table 5: Hyperparameters for Subtask 3 (ST3).

C.4 Subtask 4 (ST4) Configurations

Hyperparameter	Value
Model	microsoft/mdeberta-v3-base
Batch Size	72
Gradient Accumulation Steps	1
Effective Batch Size	72
Epochs	3
Learning Rate	5×10^{-6}
Warmup Ratio	0.06
Weight Decay	0.01
Max Sequence Length	265
Premise Buffer Size	16
Dropout Rate	0.1
Max λ_{bias}	2.0
DRO Warmup	1 Epoch
Early Stopping Patience	2
Early Stopping Min Δ	0.001
Mixed Precision (FP16)	True

Table 6: Hyperparameters for Subtask 4 (ST4).

D Detailed Ablation Results

Table 7 provides the complete set of experimental configurations and ablation studies conducted across all four subtasks.

Subtask	Configuration	Acc (↑)	F1-P (↑)	TCE (↓)	S_{rank} (↑)
ST1	Vanilla (Baseline)	100.0	–	0.0	100.0
	Bias	100.0	–	0.0	100.0
	DRO + Bias	100.0	–	0.0	100.0
	MSE + DRO + Bias†	100.0	–	0.0	100.0
	FreeLB + MSE + DRO + Bias†	100.0	–	0.0	100.0
	KL + DRO + Bias (Final)	**100.0**	–	**0.0**	**100.0**
ST2	Vanilla (Baseline)	98.95	97.89	2.13	45.99
	DRO + Bias	97.89	95.79	4.26	36.42
	Bias	98.42	96.84	3.19	40.13
	MSE + Bias†	98.42	96.84	3.19	40.13
	MSE + DRO + Bias†	97.89	96.84	2.20	45.03
	FreeLB + MSE + DRO + Bias†	99.47	98.95	1.06	57.53
	KL + DRO + Bias	99.47	98.95	1.06	57.53
	KL + DRO + Bias (Large)	**100.0**	**100.0**	**0.0**	**100.0**
ST3	Vanilla (Baseline)	91.15	–	8.29	28.23
	Bias + Hans + DRO + KL (non symbolic)	98.44	–	2.08	46.30
	Bias + Hans + DRO + MSE† $(nonsymbolic)$	98.44	–	2.08	46.30
	Bias (non-symbolic)	99.48	–	1.09	57.31
	Bias	98.96	–	1.04	57.74
	Bias + Hans + DRO + KL	99.48	–	1.04	58.05
	Bias + Hans + DRO + MSE†	99.48	–	1.04	58.05
	Bias + HANS + DRO	100.00	–	0.00	100.00
	Bias + HANS	**100.00**	–	**0.00**	**100.00**
ST4	Vanilla (Baseline, MT)	85.93	81.77	7.22	26.98
	DRO + Bias (MT)	84.89	79.68	8.26	25.50
	Bias (MT)	84.37	80.20	8.24	25.98
	KL + DRO + Bias (MT)	88.02	85.93	4.99	31.16
	KL + DRO + Bias (ET)	89.06	91.66	5.03	32.29
	ST2 Best Model (ET)	89.06	89.06	2.89	37.78
	DRO + Bias (ET)	89.06	88.02	2.30	40.35
	Bias (ET)	**89.06**	**90.62**	**2.21**	**41.43**

Table 7: System performance across all experimented subtasks and configurations. † indicates experimental loss components. F1-P denotes the F1 score for premise selection. S_{rank} for ST2-4 is the combined ranking score. For ST4, (MT) represents the original multilingual test data given by SemEval Task curators, while (ET) represents its English translation by ChatGPT (OpenAI, 2026).