ModusPonens at SemEval-2026 Task 11: Breaking the Plausibility Trap in LLMs via Conflict-Aware Ensembling

Soumyajit Roy
Setu
Bengaluru, India
roysoumyajit@icloud.com

Manav Malhotra
Chubb
Hyderabad, India
manavmalhotra173@gmail.com

Abstract

Large Language Models (LLMs) often struggle to disentangle formal logical validity from real-world plausibility, a phenomenon known as the "belief bias". This paper describes our submission to SemEval-2026 Task 11. We frame the task as a calibration problem between "System 1" (heuristic) and "System 2" (logical) thinking. Our experiments reveal that standard neuro-symbolic interventions, such as Structural Chain-of-Thought (CoT) and Nonsense Augmentation, degrade performance in low-resource regimes due to an "abstraction penalty". Instead, we propose a **Conflict-Aware Logit Ensemble**. We fine-tune two variations of Qwen-2.5-14B: a standard "Believer" model and a bias-hardened "Skeptic" model trained on oversampled conflict data. By ensembling their logits via soft-voting, we achieve a Pareto-optimal balance, reducing the Total Content Effect (TCE) to **3.21** while maintaining an overall accuracy of **94.27%**, resulting in a Combined Score of **39.09**.

1 Introduction

The ability to reason deductively, independent of semantic content, is a hallmark of robust intelligence. However, LLMs frequently fall into the "Plausibility Trap," where they incorrectly label logically Invalid arguments as Valid simply because the conclusion is factually true (e.g., "All roses are flowers... therefore some roses fade"). SemEval-2026 Task 11 (Valentino et al., 2026) challenges systems to minimise this content bias.

Our participation focuses on NLI-style classification. We hypothesised that while LLMs possess strong implicit reasoning capabilities, they lack calibration when world knowledge conflicts with logical rules. Through extensive ablation, we found that explicit structural disentanglement methods (like abstracting entities to variables) failed to generalise given the small training size (768 examples).

Our successful strategy relies on **Conflict-Aware Ensembling**. We train two distinct models: one that learns the natural data distribution, and one explicitly forced to focus on logical "traps" via weighted sampling. Combining these models allows us to filter out belief bias without sacrificing linguistic coherence.

1.1 Contributions

In summary, our contributions to SemEval-2026 Task 11 are three-fold:

1. **Identification of the Abstraction Penalty:** We provide empirical evidence that in low-resource regimes ($\sim$800 examples), standard neuro-symbolic interventions (e.g., Chain-of-Thought, variable abstraction) degrade performance for 14B-parameter models. We show that removing semantic content forces the model to struggle with variable tracking, outweighing the benefits of structural disentanglement.

2. **The "Skeptic" Training Strategy:** We introduce a targeted data-centric intervention—Conflict-Aware Weighted SFT—that fine-tunes a model specifically on the "minority class" of logical traps. We demonstrate that this successfully reverses the "Plausibility Trap," increasing accuracy on Invalid & Plausible samples from 87.2% to 95.7%.

3. **Pareto-Optimal Ensembling:** We propose a Logit-Level Soft Voting mechanism that fuses the predictions of a standard "Believer" model and a bias-hardened "Skeptic" model. This approach achieves a state-of-the-art trade-off, minimising the Total Content Effect (TCE) to 3.21 while maintaining high general accuracy, resulting in a Combined Score of 39.09.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 65–71

2 Background

The core challenge of this task is the phenomenon of Belief Bias, where a reasoner's judgement of an argument's logical validity is conflated with the believability of its conclusion. In the context of Large Language Models (LLMs), this is often framed as a conflict between "System 1" (heuristic, content-driven) and "System 2" (algorithmic, rule-driven) processing (Dasgupta et al., 2024).

The dataset for Task 11 is structured to explicitly test this calibration by dividing syllogisms into four distinct subgroups:

1. **Valid & Plausible (VP):** Consistent. The logic holds, and the conclusion matches world knowledge (e.g., "All birds have feathers; Robins are birds; Therefore, Robins have feathers").

2. **Valid & Implausible (VI):** Conflict. The logic holds, but the conclusion violates world knowledge (e.g., "All mammals are jellies; Dogs are mammals; Therefore, Dogs are jellies").

3. **Invalid & Plausible (IP):** Conflict. The "Plausibility Trap." The logic is flawed, but the conclusion is factually true (e.g., "All roses are flowers; Some flowers fade; Therefore, Some roses fade").

4. **Invalid & Implausible (II):** Consistent. The logical validity (False) and the semantic plausibility (False) align. Both the logic and the conclusion are clearly wrong, meaning heuristic and algorithmic processing lead to the same answer.

Standard LLMs, pretrained on vast corpora of natural text, exhibit a strong prior for semantic probability. When evaluating IP samples, the model's language modelling objective ($P(\text{next\_token} \mid \text{context})$) encourages it to output "True" because the conclusion statement itself has high likelihood in the pretraining distribution, overriding the structural logical violation (Lampinen et al., 2022). Our work aims to disentangle these signals by explicitly modelling the conflict.

3 Related Work

3.1 Content Effects in LLMs

The tendency of Large Language Models to rely on semantic priors rather than logical rules - often termed "belief bias" - is well-documented. Dasgupta et al. (2024) demonstrated that LLMs, like humans, struggle to decouple reasoning from world knowledge, performing significantly worse when valid arguments contradict factual beliefs. Eisape et al. (2024) further formalised this by comparing human and model performance on syllogistic reasoning, finding that models often mirror human-like non-logical heuristics. Unlike humans, however, models lack a robust "System 2" monitoring mechanism to override these priors in zero-shot settings (Bertolazzi et al., 2024).

3.2 Neuro-Symbolic and Structural Interventions

Recent approaches have attempted to mitigate these biases through architectural or prompting interventions. Chain-of-Thought (CoT) prompting has been extended to include symbolic abstractions; for instance, Ranaldi et al. (2025) proposed "quasi-symbolic" abstractions to force models to reason over variable placeholders rather than entities. Similarly, Xu et al. (2024) and Lyu et al. (2023) explored "Faithful CoT," constraining the generation process to symbolic derivation steps.

However, Maraia et al. (2026) suggest that content invariance might be better encoded in specific activation spaces rather than explicit output tokens. Our work supports this hypothesis: we find that explicit symbolic CoT (e.g., forcing the model to output "All A are B") degrades performance in 14B-parameter models, likely due to the "abstraction penalty" where removing semantic scaffolding hampers the model's tracking abilities. Instead of architectural complexity, we align with data-centric approaches (Ozeki et al., 2024) by altering the training distribution itself to penalise heuristic shortcuts.

4 System Overview

Our system is built upon **Qwen-2.5-14B-Instruct**, a decoder-only transformer chosen for its strong reasoning capabilities relative to its size, fitting within consumer GPU limits (24GB VRAM) using 4-bit quantisation.

We opted for a decoder-only model over encoder-decoder architectures (e.g., T5) due to its stronger performance in instruction-following and generative reasoning tasks, as well as compatibility with parameter-efficient fine-tuning (LoRA) under constrained GPU resources. Preliminary experiments

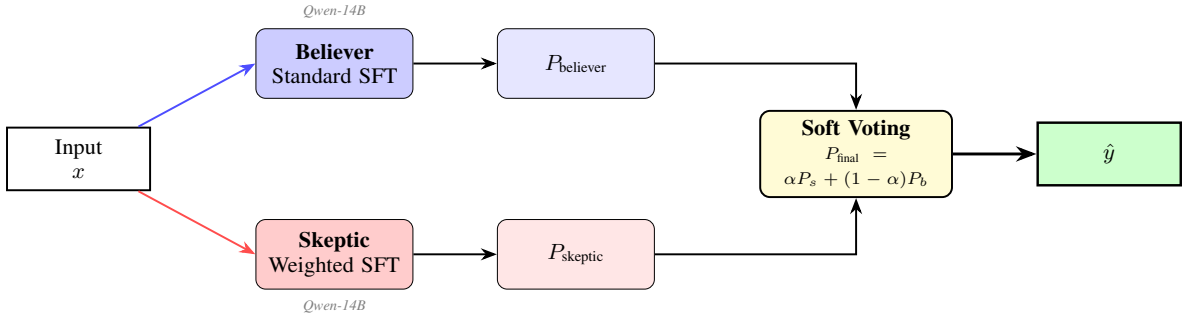

Figure 1: **The Conflict-Aware Ensemble Architecture.** The input is processed in parallel by the "Believer" (trained on natural distribution) and the "Skeptic" (trained on conflict-heavy distribution). Their logits are fused via a weighted soft-vote to produce the final prediction.

with smaller encoder-decoder models yielded lower accuracy in this low-resource regime.

4.1 The "Believer": Standard SFT

Our baseline model is fine-tuned using Standard Supervised Fine-Tuning (SFT) on the raw dataset. We use Low-Rank Adaptation (LoRA) to adapt the attention and MLP modules. This model achieves high accuracy on consistent samples (VP, II) but exhibits a high belief bias, frequently failing on the Invalid and Plausible subgroup (87.23% accuracy vs 95%+ for others). We term this model the "Believer" as it tends to trust plausible conclusions.

4.2 The "Skeptic": Conflict-Aware Weighted SFT

To counter belief bias, we identify training samples where Validity $\neq$ Plausibility. These "Conflict" samples constitute the minority of the data but represent the core challenge. We construct a weighted dataset where existing Conflict samples are oversampled by a factor of $k = 5$ through direct duplication. No new synthetic samples were generated; rather, the duplication forces the cross-entropy loss to heavily penalise heuristic shortcuts during fine-tuning. Training on this distribution forces the model to prioritise structural cues over semantic heuristics. This model, the "Skeptic," achieves significantly higher accuracy on the Invalid & Plausible trap (95.74%), effectively mitigating belief bias.

The oversampling factor of $K = 5$ was selected heuristically. In the training split, there were roughly equal numbers of conflict ($\sim$379) and consistent ($\sim$389) samples. A $5\times$ multiplier ensured that conflict samples heavily dominated the loss landscape to overcome pre-training priors, without causing catastrophic forgetting of natural language patterns. While we did not exhaustively ablate K beyond 5 due to compute constraints, optimising this ratio represents a valuable direction for future work to minimise the 'over-correction' observed in the standalone Skeptic model.

4.3 The Ensemble: Logit-Level Soft Voting

We observe that the "Believer" and "Skeptic" have uncorrelated error modes. The Skeptic effectively rejects invalid plausible statements but can become overly critical of valid arguments.

We combine them using a weighted average of their output logits:

$$P_{\text{final}}(y) = \alpha \cdot P_{\text{skeptic}}(y) + (1-\alpha) \cdot P_{\text{believer}}(y)$$

where $P(y)$ represents the softmax probability of the "True" token. We performed a discrete grid search for α over the interval [0.0,1.0] in steps of 0.1 and 0.2, evaluated on the validation set.

Furthermore, to prevent the model from generating out-of-domain text, we did not decode the output string. Instead, we performed a forward pass with the input prompt and extracted the raw logits of the last token position for the specific vocabulary IDs corresponding to the words "true" and "false". We then applied a softmax strictly over this 2-dimensional vector. This guarantees a normalised binary probability $P(True) + P(False) = 1$ and completely circumvents the issue of the model generating unexpected tokens.

4.4 Failed Approaches (Negative Results)

We attempted several other techniques which, counter-intuitively, degraded performance:

- **Structural CoT:** We fine-tuned the model to output a variable-based structure (e.g., "Structure: All A are B...") before the label.

This resulted in mode collapse (Accuracy: 58.85%), suggesting that for 14B models, the overhead of learning a new syntax distracts from the reasoning task in low-data regimes.

- **Nonsense Augmentation:** We replaced nouns with nonsense words ("All glorps are dax"). This yielded lower accuracy (88.02%) and higher bias (TCE: 14.9), indicating the model relies on "semantic scaffolding" to track variable relationships. Methodologically, we implemented this using the spaCy pipeline to extract `noun_chunks` from the premises, mapping each unique real-world entity to a fixed dictionary of anonymised pseudo-words while preserving the grammatical scaffolding. This degraded performance to 88.02%.

- **Counterfactual Logic Consistency Training (CLCT):** We hypothesised that a model should predict identical labels for a syllogism and its abstract counterpart (e.g., replacing nouns with variables A, B, C). We implemented a consistency loss term $\mathcal{L}_{\text{cons}} = \text{MSE}[1](P_{\text{orig}}, P_{\text{abs}})$ to penalise divergence. While this approach improved the robustness of the model on abstract data, it degraded performance on natural language samples (Combined Score: 23.02), likely because the abstraction process removed semantic cues that the 14B model relies on for context.

- **Preference Tuning via ORPO:** To explicitly penalise semantic priors, we applied Odds Ratio Preference Optimisation (ORPO). For every syllogism, we defined the chosen response as the correct logical boolean ("true" or "false") and the rejected response as its logical opposite. Despite running for 100 steps with a beta of 0.1, this approach severely degraded overall accuracy (77.60%) and worsened the Total Content Effect (30.56). This indicates that standard preference tuning, which is designed for conversational alignment, struggles to optimise rigid, single-token binary logic, often collapsing the model's reasoning capabilities instead of disentangling biases.

[1]MSE: Mean Squared Error

Approach	Acc	TCE ↓	Score ↑	IP Acc
Structural CoT	58.85	27.21	13.56	80.85
Nonsense Aug	88.02	14.92	23.36	68.09
ORPO	77.60	30.56	17.43	68.09
Believer	92.71	4.36	34.60	87.23
Skeptic	**95.31**	5.10	33.94	**95.74**
Ensemble	**95.31**	**3.21**	**39.09**	91.49

Table 1: Performance comparison of different approaches.

5 Experimental Setup

5.1 Data Splits

We utilised an 80/20 split of the provided training data. To guarantee an unbiased validation set, we engineered a composite stratification key by concatenating the boolean validity and plausibility labels. This explicit 4-class stratification ensured an exactly proportional distribution of the critical subgroups (VP, VI, IP, II) across both the training and validation splits.

5.2 Hyperparameters

- **Model:** unsloth/Qwen2.5-14B-Instruct-bnb-4bit
- **LoRA:** $r = 16$, $\alpha = 16$, Target Modules: `[q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj]`
- **Training:** Batch size 2, Gradient Accumulation 4, Learning Rate 2×10^{-4}, Cosine Scheduler, 1 Epoch (Standard) / 100 steps.
- **Hardware:** Single NVIDIA A40 GPU.
- **Libraries:** Unsloth, TRL, PEFT, Transformers.

5.3 Evaluation Metrics

We optimise for the official Combined Score, defined as:

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

where TCE (Total Content Effect) measures the performance gap between consistent and conflict subgroups.

6 Results

Table 1 presents our main results on the validation split.

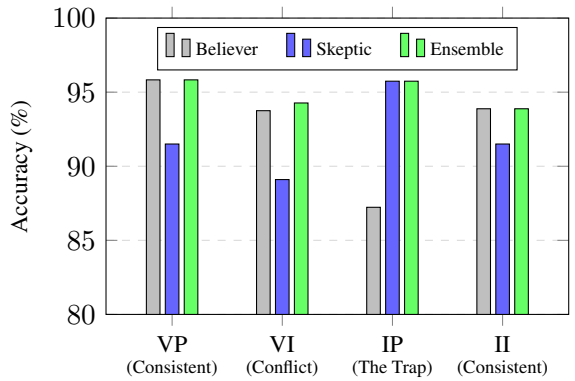

Figure 2: **Accuracy Breakdown by Logical Subgroup.** The Ensemble resolves the "Plausibility Trap" (IP) while maintaining high accuracy on consistent samples.

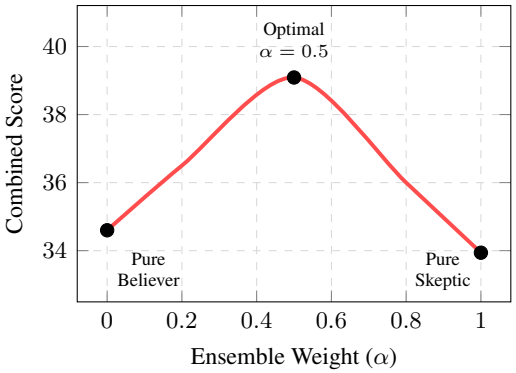

Figure 3: **Sensitivity Analysis of α.** The ensemble achieves peak performance at $\alpha = 0.5$, balancing the Believer's pattern recognition with the Skeptic's logical rigor.

6.1 The Abstraction Penalty

A key finding of our experiments is the failure of structural interventions in the low-data regime. Approaches 1 and 2 (CoT and Augmentation) aimed to force the model to reason symbolically. However, results indicate an Abstraction Penalty: for a 14B parameter model trained on only $\sim$800 examples, the overhead of learning a new, abstract syntax (e.g., "All A are B") overwhelmed the reasoning task itself.

- Structural CoT suffered from severe hallucinations, often generating invalid intermediate structures that did not match the input premises.
- Nonsense Augmentation (replacing "dogs" with "glorps") degraded performance (88.02%), suggesting the model relies on "semantic scaffolding"—familiar concepts help the model track variable relationships, even if they introduce bias.

6.2 The Skepticism Trade-off

The Weighted SFT ("Skeptic") model successfully inverted the standard failure mode. By oversampling conflict data, it achieved near-perfect performance on the Invalid & Plausible (IP) "traps" (95.74%), a massive improvement over the Baseline (87.23%). However, this came at a cost: the Skeptic became hyper-critical. It began to reject Valid & Implausible (VI) arguments at a higher rate, seemingly learning a heuristic of "if it sounds weird, it's probably valid" but misapplying it to complex valid structures. This increased the TCE score (5.10) compared to the baseline (4.36), proving that minimising belief bias alone is not sufficient if it introduces a new "skepticism bias."

For example, given a VI sample such as *'All mammals are jellies. Dogs are mammals. Therefore, dogs are jellies'*, the Baseline model correctly identifies the valid logical structure. However, the Skeptic model over-corrects and predicts 'False,' treating the bizarre semantic content as a proxy for invalidity.

6.3 Pareto Optimality via Ensembling

The Logit Ensemble ($\alpha = 0.5$) represents the Pareto-optimal point between these two extremes.

- It retains the Baseline's high precision on consistent data (System 1 strengths).
- It integrates the Skeptic's robustness on logical traps (System 2 strengths).
- **Mechanism:** On ambiguous samples where the Baseline assigns high probability due to plausibility (e.g., $P_{\text{believer}}(\text{True}) = 0.9$), the Skeptic acts as a veto (e.g., $P_{\text{skeptic}}(\text{True}) = 0.1$). The ensemble averages these to $\sim$0.5, pushing the decision closer to the decision boundary and allowing the model's uncertainty to be reflected.
- This smoothing effect resulted in the lowest content bias (TCE 3.21) and the highest overall Combined Score (39.09), demonstrating that soft-voting effectively calibrates the model's confidence against its semantic priors.
- In cases of extreme uncertainty where the ensemble output approaches an exact 0.5 split,

our system defaults to strict thresholding ($> 0.5 \rightarrow$ True). Consequently, the model leans slightly towards 'False' (Invalid) when entirely uncertain, demanding a higher burden of proof to validate an argument.

7 Conclusion

Our experiments demonstrate that in the context of relatively low-resource parameter models, "data-centric" interventions (conflict-aware sampling) proved more effective and stable than "architectural" interventions (CoT, Neuro-symbolic). We identified that models suffer from an abstraction penalty in low-resource settings. Our winning system, a Logit Ensemble of a standard and a conflict-weighted model, effectively neutralises belief bias, offering a robust baseline for formal reasoning in LLMs.

Limitations

While our Conflict-Aware Ensemble mitigates content effects, it introduces several computational and methodological limitations.

Computational Cost: Our approach requires running two 14B-parameter models (the "Believer" and "Skeptic") in parallel, doubling memory and compute requirements compared to a single-model baseline. While acceptable for a shared task submission, this is inefficient for deployment. Future work could explore model merging techniques (e.g., Ties-Merging or Task Arithmetic) to combine the LoRA adapters into a single model while preserving ensemble benefits.

Scope of Logical Forms: Our analysis is limited to classical Aristotelian syllogisms. It remains unclear whether the Skeptic model's robustness generalises to other logical domains such as propositional or modal logic. The anti-bias behaviour may reflect structural reasoning improvements or simply dataset-specific phrasing patterns.

Over-Correction: The Skeptic model sometimes over-corrects by rejecting valid but implausible arguments (e.g., "All trees are dogs"). This suggests that suppressing System 1 heuristics is incomplete. Future work could apply intervention-based interpretability to identify and dampen attention heads responsible for retrieving world knowledge during logical inference.

References

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. *Preprint*, arXiv:2406.11341.

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.

Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. A systematic comparison of syllogistic reasoning in humans and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *Preprint*, arXiv:2301.13379.

Gabriele Maraia, Marco Valentino, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2026. Abstract activation spaces for content-invariant reasoning in large language models. *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*, pages 16063–16077, Bangkok, Thailand. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 17222–17240. Association for Computational Linguistics.

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)*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. *Preprint*, arXiv:2405.18357.