Disentangling Content and Formal Reasoning in Large Language Models

Chaitanya Shah Shlok Sand Inesh Dheer

1 Introduction

Large Language Models (LLMs) are good at answering questions, but sometimes they mix up **logic** with **world knowledge**. For example, if an argument sounds true in real life, they may think it is logically valid even when it is not. This problem is called the *content effect*. SemEval-2026 Task 11 looks at this problem by testing how well LLMs can handle logic problems (syllogisms) in different languages, without being influenced by how believable the statements sound.

This project aims to build a system that can judge whether an argument is logically correct, without relying on how realistic it seems, and that works well across many languages.

2 Motivation

Recent research has shown that LLMs:

- Overestimate the validity of arguments aligned with common knowledge.
- Underestimate the validity of logically sound but implausible arguments.
- Exhibit biases based on argument content and language.

This *content effect* stems from the entanglement between reasoning and content learned during pre training. Addressing it is critical for:

- Improving the reliability of reasoning systems in high-stakes domains.
- Enhancing cross-lingual generalization in multilingual contexts.
- Reducing susceptibility to spurious correlations and decision-making biases.

3 Task Overview

SemEval-2026 Task 11 consists of four subtasks:

Subtask 1: Syllogistic Reasoning in English: Predict the formal validity of syllogisms in English.

Subtask 2: Syllogistic Reasoning with Irrelevant Premises in English: Predict validity while filtering out noisy premises.

Subtask 3: Multilingual Syllogistic Reasoning: Extend reasoning to multiple languages and evaluate cross-lingual robustness.

Subtask 4: Multilingual Syllogistic Reasoning with Irrelevant Premises: Combine multilingual reasoning with noisy premise filtering.

The training set will be exclusively in English to simulate a low resource multilingual transfer scenario. Arguments are annotated with both `validity` and `plausibility`, but only validity should be predicted.

4 Proposed Methodology

Our approach will combine **natively multilingual open-weight models** with **reasoning-specific adaptation techniques** to address the content effect. The methodology includes:

4.1 Base Model Selection

We will experiment with multilingual transformer-based LLMs such as `mBERT`, `XLM-RoBERTa`, and `mT5`, chosen for their strong cross-lingual transfer capabilities.

4.2 Representation Disentanglement

When a model reasons about an argument, its internal representations often mix two things: (1) the logical structure of the argument, and (2) the real-world meaning or familiarity of the statements. This mixing leads to the *content effect*, where models may rely on how realistic something sounds rather than on whether it is logically valid.

To address this, we will focus on techniques that help the model clearly separate **logical form** from **content plausibility**:

- **Neuro symbolic integration**: We will combine symbolic reasoning tools such as syllogism parsers that break down arguments into formal logic rules with neural network encoders that capture language patterns. This way, the neural model is guided by strict logical structure, reducing the chance that it is swayed by content.
- **Activation steering**: Building on the method of Valentino et al. (2025), we will identify specific neurons or attention heads in the model that strongly respond to content plausibility. During inference, we will adjust (or "steer") their activations to reduce this bias, helping the model focus more on the formal reasoning path.
- **Quasi symbolic abstractions**: Inspired by Ranaldi et al. (2025), we will design structured "chain-of-thought" templates that highlight the logical relationships between premises and conclusions, while minimizing semantic interpretation. These templates will encourage the model to treat arguments as abstract logical forms rather than as meaningful real-world claims.

Overall, these techniques aim to make the model's internal reasoning process cleaner, with a clearer boundary between what is logically necessary and what simply "sounds true."

4.3 Training Strategy

1. Supervised fine-tuning on English training data.
2. Data augmentation with synthetic syllogisms in multiple languages.
3. Contrastive learning objectives to explicitly penalize plausibility-based reasoning.

4.4 Evaluation

We will follow the competition metrics:

- **Accuracy**
- **Content Effect** (intra- and cross-plausibility)
- **F1 score** for premise selection (Subtasks 2 & 4)
- **Multilingual Content Effect**

5 Expected Outcomes

- A reasoning model that achieves high validity accuracy while minimizing content effect.
- Insights into cross-lingual transfer of formal reasoning.
- Open-source release of model weights, code, and multilingual synthetic reasoning datasets.

6 References

References

[1] Valentino, M., Kim, G., Dalal, D., Zhao, Z., & Freitas, A. (2025). Mitigating Content Effects on Reasoning in Language Models through Fine-Grained Activation Steering. *arXiv preprint arXiv:2505.12189.*

[2] Ranaldi, L., Valentino, M., and Freitas, A. (2025). Improving chain-of-thought reasoning via quasi-symbolic abstractions. *ACL 2025.*

[3] Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. (2022). Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051.*