Disentangling Content and Formal Reasoning in Large Language Models

Inesh Dheer
IIIT Hyderabad
inesh.dheer@research.iiit.ac.in

Chaitanya Shah
IIIT Hyderabad
chaitanya.shah@students.iiit.ac.in

Shlok Sand
IIIT Hyderabad
shlok.sand@research.iiit.ac.in

Abstract

Large Language Models (LLMs) can write, answer, and code, but they often mistake factual knowledge for logical reasoning. This weakness, known as the content effect, makes them unreliable because they may accept flawed arguments if the conclusion "sounds true." This project directly addresses this issue by building a system that judges a logical argument based only on its structure, not its content. The goal is to develop a model that can ignore whether statements are believable or absurd and apply these logic skills across multiple languages, proving it has learned to reason abstractly rather than just memorize patterns. By doing this, we hope to contribute to building more robust and reliable AI.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in writing, answering questions, and generating code. However, a significant limitation is their tendency to conflate factual knowledge with logical reasoning. Like a student who memorizes without understanding, they may accept flawed arguments if the conclusion seems plausible. This vulnerability, known as the content effect, makes LLMs unreliable for tasks requiring rigorous logic.

Our project directly addresses this critical gap by building a system that:

- Judges a logical argument based only on its structure, not its content.
- Ignores whether the statements sound believable or absurd in the real world.
- Applies these logic skills to many different languages, proving it has learned to reason, not just memorize.

By focusing on these objectives, we aim to contribute to the development of more robust and trustworthy AI systems.

2 Background

The "content effect" is not just an academic issue; it makes LLMs unreliable in practical applications. A model might accept a flawed legal or scientific argument simply because the conclusion seems fair, with potentially severe consequences. Our work builds on prior research, inspired by key ideas from recent studies.

Dasgupta et al. (2022) demonstrated that LLMs, like humans, are biased by the plausibility of an argument's content. This foundational work proves the existence and severity of the problem.

Building on this, Valentino et al. (2025) identified neural components within models that overreact to content. They showed that these components can be "steered" during inference to focus more on logical structure, a technique known as activation steering.

Furthermore, Ranaldi et al. (2025) introduced the use of logic templates, or quasi-symbolic abstractions, to improve reasoning. This method involves converting problems into an abstract form (e.g., "All X are Y") to strip away distracting details and highlight the underlying logical structure.

Our project will combine these cutting-edge techniques to create a system that is more than the sum of its parts, a model that can reason clearly and consistently, regardless of the topic.

3 Project Goals

We have defined three primary goals for our project:

1. **High Accuracy:** The model must reliably determine whether an argument is logically valid or invalid.
2. **Bias Resistance:** Accuracy should hold even for arguments that sound absurd. This is measured by a low "content effect" score.

3. **Cross-Language Generalization:** Though trained only on English, the model should apply logic consistently in other languages, proving it learns abstract principles, not just surface-level patterns.

4 The Task and Data

Our project directly follows the structure of the SemEval-2026 Task 11, focusing on logical reasoning in syllogisms.

4.1 The Dataset

The data consists of logical puzzles called syllogisms. The training set is entirely in English. Each puzzle has two important labels:

- `validity`: A 'true' or 'false' label indicating if the conclusion logically follows from the premises. **This is what we must predict.**
- `plausibility`: A 'true' or 'false' label indicating if the statements align with our knowledge of the real world. **This is the information that tries to trick the model.**

4.2 The Four Subtasks

The competition is broken down into four challenges of increasing difficulty:

1. **Syllogistic Reasoning in English:** The baseline task. Can the model correctly assess the validity of English syllogisms?
2. **Reasoning with Irrelevant Premises in English:** A tougher version where distracting, useless sentences are added to the puzzle. The model must learn to identify and ignore them.
3. **Multilingual Syllogistic Reasoning:** The model is tested on new languages. This measures its ability to transfer its reasoning skills.
4. **Multilingual Reasoning with Irrelevant Premises:** The ultimate challenge, combining the multilingual and irrelevant information tests.

5 Methodology: Knowledge Distillation with Symbolic Chain-of-Thought

Our novel approach employs a combination of **Knowledge Distillation** augmented with **Symbolic Chain-of-Thought (S-CoT)** reasoning to produce a compact, language-capable specialist for syllogistic and formal logical reasoning. The distillation pipeline is organized around two models:

- **Teacher model:** `deepseek-r1:8b`. The teacher is responsible for generating high-fidelity symbolic reasoning traces and labeled training examples that explicitly expose the underlying logical structure of arguments.
- **Student model:** `gemma3:4b-it-q4_K_M`. The student is a smaller, multilingual specialist that is fine-tuned to acquire symbolic reasoning patterns from the teacher and to generalize them across languages.

6 Technical Plan: Three-Part Strategy

6.1 Step 1: Teacher–Student Setup and Rationale

We adopt a two-model distillation paradigm involving a high-capacity teacher and a compact multilingual student. The teacher is fine-tuned on the annotated dataset to generate high-quality symbolic chain-of-thought (S-CoT) traces and gold labels that explicitly expose the logical structure of each instance. These teacher-produced S-CoT examples form the distilled corpus used for student training. The student, a smaller multilingual specialist, is then fine-tuned to internalize the teacher's symbolic reasoning patterns. By learning abstract symbolic transformations rather than language-specific heuristics, the student can generalize these reasoning processes across languages with minimal additional adaptation.

6.2 Step 2: Implement advanced reasoning mechanisms, towards multi-lingual generalization

To focus model learning on logical structure rather than surface plausibility, we combine three complementary techniques:

1. **Neuro-symbolic integration:** For each training instance we extract an explicit logical form (e.g., canonicalized statements like "All A are B; Some B are C; therefore..."). The logical form is provided alongside the original input so the model learns to map between natural language and a compact symbolic representation.
2. **Activation steering:** We empirically identify neurons or subspaces correlated with spurious plausibility heuristics and attenuate their influence during fine-tuning (via gradient surgery, targeted regularization, or controlled weight

updates). This reduces reliance on shallow heuristics and improves robustness to distractor premises.

3. **Symbolic Chain-of-Thought (S-CoT):** Training prompts and teacher outputs explicitly include stepwise symbolic traces that first abstract entities to variables and then perform structural inferences. Requiring the student to reproduce these traces forces intermediate computation that emphasizes form over lexical idiosyncrasies which will in turn lead to multi-lingual generalization.

7 How We Will Measure Success

We will follow the official evaluation metrics of the SemEval task, which are designed to reward models that are both accurate and unbiased. The primary ranking will be based on the ratio of Accuracy to Content Effect.

- **Accuracy:** The percentage of puzzles solved correctly.
- **Content Effect:** A measure of the performance drop between plausible and implausible cases. A lower score is better, indicating robustness against bias.
- **F1 Score:** Used for tasks involving irrelevant premises to evaluate how well the model identifies the correct sentences needed for the logical judgment.

A high final score will indicate that the model is not just accurate but also a consistent and robust reasoner.

8 Project Timeline

Our project will run from August 10 to November 1, divided into five phases:

1. **Phase 1: Foundation and Planning (Weeks 1–2, Aug 10–23).**
   We will set up our coding environment, study key research papers in depth, and analyze the sample dataset to understand its nuances.
2. **Phase 2: Building the Baseline (Weeks 3–4, Aug 24–Sep 6).**
   We will train a simple model using XLM-RoBERTa, establishing a baseline score to improve upon.
3. **Phase 3: Developing Advanced Methods (Weeks 5–8, Sep 7–Oct 4).**
   This is the main research phase, where we will implement and test three core techniques: neuro-symbolic integration, activation steering, and quasi-symbolic prompting.
4. **Phase 4: Evaluation and Analysis (Weeks 9–10, Oct 5–18).**
   We will evaluate all models on the test data, comparing their performance with a focus on accuracy and the content effect score.
5. **Phase 5: Finalization (Weeks 11–12, Oct 19–Nov 1).**
   We will prepare our final project report, clean up the codebase, and complete the final submission.

References

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, James L. McClelland, and F. Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

L. Ranaldi, M. Valentino, and A. Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Fictional entry for proposal.

Marco Valentino, Geewook Kim, Danish Dalal, Zhaozhen Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *arXiv preprint arXiv:2505.12189*.