AICOE-Tredence at SemEval-2026 Task 11: Mitigating Content Bias in Syllogisms via Symbolic Logic-Language Decoupling

Rakshith R
AICOE, Tredence
Bengaluru, India
rakshith.r@tredence.com

Ankush Chopra
AICOE, Tredence
Bengaluru, India
ankush.chopra@tredence.com

Abstract

Content bias remains a key limitation of large language models (LLMs), which often conflate formal logical validity with real-world plausibility. SemEval-2026 Task 11 examines this challenge through multilingual syllogistic reasoning, requiring models to judge validity independently of content. We propose a structure-first reasoning paradigm that abstracts natural language syllogisms into Aristotelian logical forms. By mapping arguments to mood–figure representations and classifying validity in this symbolic space, our approach removes semantic content from the reasoning process. On the private test sets of Subtasks 1 and 3, our method achieves a perfect combined score, with 100% validity accuracy and zero content bias in both English and multilingual settings using Gemini-3 Pro Preview. We also explore transferring this paradigm to smaller models via structural supervision, finding that distilled systems retain high accuracy with minimal bias. These results suggest that explicitly separating logical form from linguistic content is a promising direction for bias-resilient and cross-lingually robust reasoning in LLMs.

1 Introduction

Large language models (LLMs) show strong reasoning abilities but remain vulnerable to content effects, where semantic plausibility influences judgments of logical validity (Dasgupta et al., 2024). As a result, models may accept invalid yet believable arguments or reject valid but counterintuitive ones, violating the principle that logical validity should depend only on structure.

SemEval-2026 Task 11 (Valentino et al., 2026) evaluates this challenge through multilingual syllogistic reasoning, measuring both logical accuracy and content bias. We focus on Subtasks 1 and 3, which assess validity classification in English and across multiple languages.

We address this problem using a structure-first approach that abstracts syllogisms into Aristotelian logical forms. Arguments are mapped into mood–figure representations and classified in a symbolic space, making reasoning invariant to plausibility and language.

We implement this method using Gemini-3 Pro Preview for structural extraction and transfer it to smaller models via distillation. Our approach achieves perfect logical robustness on SemEval-2026 Task 11, eliminating content bias in both English and multilingual settings. We further distill the structural signal into Qwen-3 14B (Yang et al., 2025), demonstrating that the method generalizes beyond a single backbone.

2 Related Work

Prior work shows that large language models exhibit content effects in formal reasoning, where semantic plausibility influences logical judgments (Dasgupta et al., 2024; Mondorf and Plank, 2024; Wysocka et al., 2025). Models often perform better when arguments align with real-world knowledge and struggle with valid but counterintuitive conclusions (Bertolazzi et al., 2024).

Early evidence from (Dasgupta et al., 2024) showed higher accuracy on semantically plausible yet logically invalid syllogisms, indicating a preference for material reasoning over formal logic. Later studies reported performance drops on logically valid but counterintuitive cases (Bertolazzi et al., 2024). While prompting strategies such as chain-of-thought can improve reasoning (Wei et al., 2023), content-driven biases often persist.

3 Dataset Overview

3.1 Subtask 1

The Subtask 1 training set consists of 960 English syllogisms annotated for both logical validity and plausibility. The dataset is balanced with 480 valid

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1802–1816

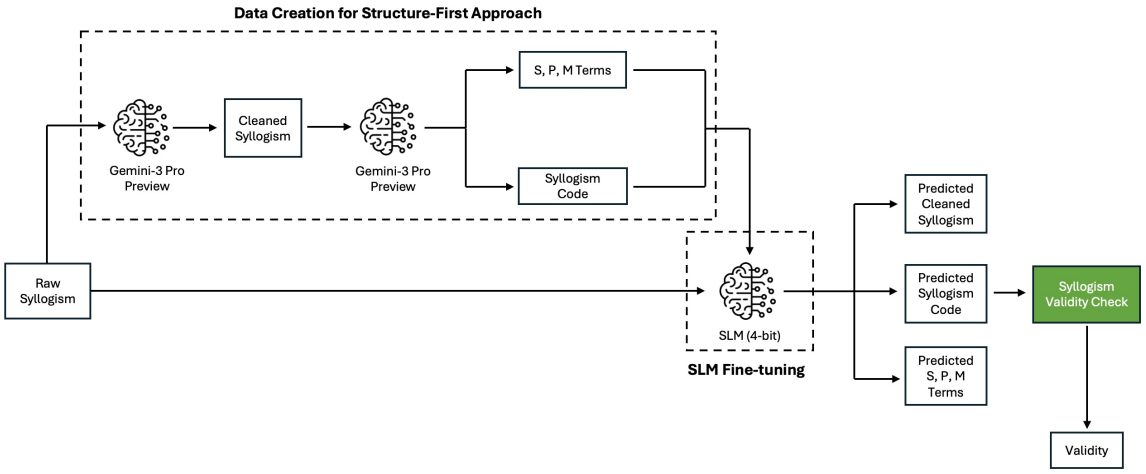

Figure 1: This figure illustrates the workflow of our Structure First Approach

and 480 invalid instances, and plausibility labels are similarly distributed (486 implausible vs. 474 plausible). The joint distribution across validity and plausibility is approximately uniform, ensuring coverage across all content-bias conditions. Syllogistic validity in this dataset follows classical Aristotelian logic, under which only 24 of the 256 possible mood–figure combinations are considered valid (Zhang and Qiao, 2009)

The private test set contains 191 English syllogisms. Validity and plausibility labels are not provided for this split, and evaluation is performed via the official leaderboard.

Throughout this work, we distinguish between different forms of syllogisms. We refer to the original inputs provided by the dataset as raw syllogisms (e.g., "Not a single bird is a mammal. All sparrows are birds. Therefore, no sparrows are mammals."). After applying normalization steps such as removing linguistic noise and canonicalizing quantifiers, we obtain cleaned syllogisms (e.g., "No birds are mammals. All sparrows are birds. Therefore, no sparrows are mammals."). These cleaned versions preserve logical structure while eliminating surface variability.

3.2 Subtask 3

Subtask 3 focuses on multilingual syllogistic reasoning across 11 languages: German, Spanish, French, Italian, Dutch, Portuguese, Russian, Chinese, Swahili, Bengali, and Telugu. Unlike Subtask 1, multilingual training data is not provided. To address this, we derive multilingual training datasets from the Subtask 1 English corpus using augmentation techniques, described in the following sections.

The private test set contains 192 syllogisms spanning the same 11 languages. Validity and plausibility labels are not provided for this split, and evaluation is conducted via the official leaderboard.

4 Experimentations

4.1 Subtask 1

For Subtask 1, we evaluate multiple modeling strategies, including direct validity prediction and our proposed structure-first method. Each experimental setup is conducted on a different dataset derived from the original 960 samples through targeted augmentation techniques. In the following sections, we describe each method along with the corresponding data used and the resulting performance.

Evaluation uses validity accuracy and content bias, with the combined score balancing accuracy against bias.

4.1.1 Zero-shot LLM Benchmarking

To establish a performance baseline, we evaluate several proprietary and open-weight LLMs on the private test set without fine-tuning. We test GPT-5, Gemini-3 Pro Preview, and Gemini-3 Flash Preview, alongside open-weight models like Gemma-3 27B (Team et al., 2025), and Qwen-3 14B. Models are explicitly prompted to prioritize formal logical validity over semantic plausibility.

The results in Table 1 show that Gemini-3 Pro Preview is the top-performing model in the zero-shot setting.

Strategy	Model	Holdout Test Set			Private Test Set		
		Comb	Val Acc	Bias	Comb	Val Acc	Bias
Zero-Shot Benchmarking	GPT-5	–	–	–	19.61	80.10	20.83
	Gemini-3 Pro Preview	–	–	–	**22.11**	84.29	15.62
	Gemini-3 Flash Preview	–	–	–	21.58	**84.82**	17.70
	Qwen-3 14B	–	–	–	21.64	79.58	**13.54**
	Gemma-3 27B	–	–	–	10.19	50.26	50.00
Direct Prediction	Gemma-3 12B (4-bit)	38.92	97.20	3.10	40.54	98.43	3.17
	Qwen-3 14B (4-bit)	**55.81**	**98.70**	**1.00**	**58.05**	**99.48**	**1.04**
Direct + Aug	Gemma-3 12B (4-bit)	–	97.82	–	57.68	99.48	1.06
	Qwen-3 14B (4-bit)	–	**98.56**	–	**58.05**	**99.48**	**1.04**
Structure-First	Gemini-3 Pro	**100.0**	**100.0**	**0.0**	**100.0**	**100.0**	**0.0**
	Qwen-3 14B (4-bit)	**100.0**	**100.0**	**0.0**	**100.0**	**100.0**	**0.0**

Table 1: Comparison of modeling strategies on holdout and private test sets for Subtask 1. **Direct**, **Direct + Aug**, and **Structure-first** correspond to **Direct Validity Prediction**, **Direct Validity Prediction with Data Augmentation**, and **Structure-First Approach** respectively.

Strategy	Model	Holdout Test Set			Private Test Set		
		Comb	Val Acc	Bias	Comb	Val Acc	Bias
Zero-Shot Benchmarking	GPT-5	–	–	–	19.16	79.17	21.87
	Gemini-3 Pro Preview	–	–	–	**22.06**	**85.42**	16.66
	Gemini-3 Flash Preview	–	–	–	21.92	84.90	16.66
	Qwen-3 14B	–	–	–	21.61	76.56	**11.70**
	Gemma-3 27B	–	–	–	10.13	50.00	50.00
Direct + Aug	Gemma-3 12B	–	97.21	–	38.50	95.60	3.40
	Qwen-3 14B (4-bit, r=32)	–	97.89	–	40.08	96.88	**3.12**
	Qwen-3 14B (4-bit, r=64)	–	**98.12**	–	**40.73**	**98.44**	**3.12**
Structure-First	Gemini-3 Pro Preview (V1)	–	–	–	36.86	97.40	4.17
	Gemini-3 Pro Preview (V2)	–	–	–	**100.0**	**100.0**	**0.0**
	Qwen-3 14B (4-bit, r=32)	39.90	98.12	3.30	30.69	91.67	6.29

Table 2: Comparison of modeling strategies on holdout and private test sets for Subtask 3

4.1.2 Direct Validity Prediction

In this setting, validity is predicted directly from the raw syllogism with instructions to ignore plausibility. No augmentation is applied, and the original 960 samples are used as-is. The data is split into 768 training, 96 validation, and 96 holdout samples.

We fine-tune Qwen-3 14B and Gemma-3 12B using 4-bit quantization with the Unsloth framework (Daniel Han and team, 2023). Training is conducted using LoRA ($r = 32$, $\alpha = 32$), the AdamW optimizer, a learning rate of 1×10^{-4}, and a batch size of 64. Models are trained on the training split, validated on the validation split, and evaluated on the holdout set.

Qwen-3 14B outperforms Gemma-3 12B on both holdout and private test sets. Detailed results are shown in Table 1.

4.1.3 Direct Validity Prediction with Data Augmentation

To improve diversity, we perform augmentation on the original set of raw syllogisms using Gemma-3 27B.

Data Augmentation. We use two strategies. First, structure-level augmentation generates label-inverted variants by modifying the logical form while preserving surface patterns. Specifically, for every valid syllogism we generate two invalid variants, and for every invalid syllogism we generate two valid variants. This expands the original dataset of 960 raw syllogisms (480 valid, 480 invalid) to a total of 2,880. Second, entity replacement creates two variants per syllogism by substituting entities while preserving structure, resulting in 8,640 raw syllogisms (6,912 train, 864 validation, 864 holdout).

We fine-tune Gemma-3 12B and Qwen-3 14B using the same quantized setup described earlier. Be-

cause plausibility annotations were not synthesized during the data augmentation phase, the holdout split lacks these labels. Consequently, evaluation on this split is restricted to validity accuracy.

Qwen-3 14B outperforms Gemma-3 12B on holdout accuracy and private test results under this strategy. Detailed scores are shown in Table 1.

4.1.4 Structure-First Approach

Our approach explicitly decouples logical structure from linguistic content. We map each syllogism to its Aristotelian form by identifying its categorical types. The categorical types are A (universal affirmative: "All S are P"), E (universal negative: "No S are P"), I (particular affirmative: "Some S are P"), and O (particular negative: "Some S are not P").

The sequence of these types across the two premises and conclusion determines the syllogistic mood (e.g., AAA). The figure is determined by the relative positions of the subject (S), predicate (P), and middle (M) terms across the premises. Together, the mood and figure define a form such as AAA-1, which is then validated against the 24 valid Aristotelian syllogisms.

The pipeline involves two steps. First, each raw syllogism is normalized to produce a cleaned syllogism. Second, the model extracts the S, P, and M terms and generates the syllogistic code. Validity is determined solely from this symbolic representation, making the process invariant to semantic content.

Using Gemini-3 Pro Preview, this method achieved a perfect combined score of 100 on both holdout and private test sets. The prompt templates used in this pipeline are provided in the Appendix A.1. Hence, we use this model to generate training data for distillation into smaller models.

Data Creation During initial data validation, we discard 4 of the original 960 syllogisms because they were incorrectly formulated and lacked a middle term. Using the remaining 956 samples, we construct a structurally supervised dataset to enable distillation. We first extract the syllogistic mood by classifying each of the three statements and combining them into a unified mood (e.g., AEA). The prompt template used for this sentence-level mood classification is provided in Appendix A.2. The syllogistic figure is then computed by extracting S, P, and M terms and determining their positional relationships using deterministic rules implemented in Python to avoid model hallucinations. In addition, we include cleaned syllogisms as part of the supervision targets.

Each training instance consists of a raw syllogism and a target containing the cleaned syllogism, extracted S, P, and M terms, and the syllogism code. Although validity depends only on the syllogism code, we supervise intermediate outputs such as cleaned syllogisms and extracted terms. This follows the intuition that predicting intermediate structure improves robustness and accuracy.

Using this dataset, we fine-tune Qwen-3 14B with the same quantized fine-tuning configuration described earlier. The end-to-end structure-first fine-tuning prompt used to train this model is provided in Appendix A.3. This model achieves a combined score of 100 on both evaluation sets. Detailed results are shown in Table 1. A detailed overview of this approach is shown in Figure 1.

4.2 Subtask 3

For Subtask 3, we evaluate multiple modeling strategies, including direct validity prediction and our proposed structure-first method. As multilingual training data are not provided, we construct multilingual datasets by augmenting the original English syllogisms of Subtask 1. Each experimental setup is conducted on these derived multilingual datasets. In the following sections, we describe each method along with the corresponding data used and the resulting performance.

Performance is evaluated using the same metrics as Subtask 1.

4.2.1 Zero-shot LLM Benchmarking

We perform zero-shot benchmarking on the private test set. We evaluate proprietary models (GPT-5, Gemini-3 Pro Preview and Flash Preview) and open-weight models (Gemma-3 27B and Qwen-3 14B), instructing them to focus on logical validity rather than plausibility. Gemini-3 Pro Preview achieves the best performance. Results are shown in Table 2.

4.2.2 Direct Validity Prediction

In this setting, models directly predict logical validity from multilingual raw syllogisms without explicit structural abstraction.

Data Creation. For training, we construct a larger multilingual corpus by translating the augmented English raw syllogisms from Subtask 1, specifically the Direct Validity Prediction with Data

Lang.	Raw Syllogism & English Translation	GT	Pred.	Error Analysis
RU	Дело не в том, что газированные напитки — это вода. Вся существующая вода является напитком. Некоторые из существующих напитков не являются газировкой. [Trans: The point is not that carbonated drinks are water. All existing water is a drink. Some existing drinks are not carbonated.]	EAO-4	AOO-3	The model missed the conversational negation in P1 ("The point is not that..."), incorrectly predicting 'A' instead of 'E'.
SW	Hakuna televisheni ambazo sio hisia. Ni ukweli kwamba kila mhemko ni kiumbe mdogo. Hii inasababisha hitimisho kwamba mambo kadhaa madogo ni televisheni. [Trans: There are no televisions which are not feelings. It is true that every emotion is a small creature. This leads to the conclusion that some small things are televisions.]	AAI-4	EAI-4	The phrase "Hakuna... ambazo sio" (There are no... which are not) means "All" (A). The model saw the negative keyword and lazily predicted 'E'.
SW	Baadhi ya ndege ni wanyama kipenzi. Ndege wote wana manyoya. Kwa hivyo, baadhi ya wanyama wenye manyoya ni wanyama kipenzi. [Trans: Some birds are pets. All birds have feathers. Therefore, some feathered animals are pets.]	IAI-3	IAI-4	The model got the Mood (IAI) perfectly correct, but misidentified the position of the Middle Term ("birds") due to native syntax, shifting Figure 3 to 4.

Table 3: Selected error cases from Russian and Swahili in the Subtask 3 private test set from Qwen3 14b finetuning. The **GT** represents the Ground Truth syllogism code, while **Pred.** represents the predicted syllogism code.

Augmentation setup, into 11 target languages using Gemma-3 27B. Starting from 8,640 English samples, this results in 95,040 multilingual raw syllogisms (76,032 train, 9,504 validation, 9,504 holdout). Since the holdout split lacks plausibility annotations, content bias cannot be computed and evaluation is limited to validity accuracy.

We fine-tune Gemma-3 12B (4-bit) and Qwen-3 14B (4-bit) using the Unsloth framework on this dataset. Gemma-3 12B is trained with LoRA rank $r = 32$ and $\alpha = 32$, while Qwen-3 14B is evaluated under two configurations ($r = \alpha = 32$ and $r = \alpha = 64$). All configurations use the AdamW optimizer and a batch size of 32.

Among these, Qwen-3 14B with $r = \alpha = 64$ performs best within this strategy. Detailed results are shown in Table 2.

4.2.3 Structure-First Approach

Data Construction. We construct a multilingual dataset using 956 raw English syllogisms as the source, translated into 11 languages using Gemini-3 Pro Preview. The translation prompt used in this step is provided in Appendix A.5.

We extend the structure-first pipeline to the multilingual setting by abstracting each syllogism into a language-invariant logical form before validity classification.

We evaluate two variants using Gemini-3 Pro Preview. In the first variant, each multilingual raw syllogism is translated into an English raw syllogism, then normalized into a cleaned syllogism, and finally processed to extract S, P, M and the syllogistic code. In the second variant, the model directly produces cleaned syllogisms and extracts S, P, M and the final code in the native language. Results on the private test set (Table 2) show that the second variant achieves a perfect combined score of 100, and we use this approach for distillation. The native-language structure extraction prompt for the second variant is provided in Appendix A.4.

Using this pipeline, we construct a structurally supervised multilingual dataset containing cleaned syllogisms, native-language S, P, M terms, and syllogistic codes, resulting in 10,516 instances (8,412 train, 1,052 validation, 1,052 holdout). The prompt template used to generate this dataset is provided in Appendix A.6.

We fine-tune Qwen-3 14B (4-bit) on this dataset using LoRA ($r = 64$, $\alpha = 64$) using a batch size of 32 and AdamW optimizer. The prompt template used for fine-tuning is provided in Appendix A.7. Detailed results are reported in Table 2.

5 Discussion

We begin by examining zero-shot LLM behavior across both subtasks. Across proprietary models such as GPT-5 and Gemini-3 variants and open-weight models like Gemma-3 27B and Qwen-3 14B, content bias persists (Tables 1, 2). Stronger models achieve higher accuracy but do not eliminate bias, suggesting scale alone does not resolve content effects. Even after fine-tuning Gemma-3 12B and Qwen-3 14B with instructions to ignore plausibility, measurable bias remains (Tables 1, 2).

Data augmentation in Subtask 1 improves performance for Gemma-3 12B, but has limited impact on Qwen-3 14B (Table 1). The improvements are mainly in accuracy rather than bias reduction. The persistence of bias even after augmentation suggests that increasing data diversity alone is not sufficient to separate logical reasoning from linguistic content.

In contrast, the structure-first pipeline, implemented using Gemini-3 Pro Preview, demonstrates that explicitly modeling logical form can eliminate content bias entirely (Tables 1, 2). By extracting mood–figure representations and performing validity checks in symbolic space, the approach avoids reliance on surface semantics. This leads to perfect robustness in Subtask 1 and extends effectively to Subtask 3 when applied directly in the native language setting.

Distilling this structural supervision into smaller models shows promising but nuanced results. The distilled Qwen-3 14B model retains zero bias in English, confirming that structural reasoning signals transfer well (Table 1). However, some bias reappears in the multilingual setting (Table 2). This is likely due to translation noise and distribution shifts between the holdout and private test sets, suggesting that multilingual reasoning pipelines are more sensitive to upstream noise.

This sensitivity is clear in the results on the private test set. In eight out of eleven languages, the model performed very well, with accuracy ranging from 16 out of 17 to a perfect 18 out of 18. However, the model took a larger hit in specific languages, correctly predicting 13 out of 17 cases in Russian, and dropping to 12 out of 17 in Swahili.

These errors happen for two main reasons: Mood mismatches and Figure shifts. Mood errors occur when complex sentences hide the logic of words like "all," "some," or "not." Figure shifts happen when the sentence structure makes it hard to find the correct position of the Middle Term. We show examples of these errors in Table 3. These results highlight the ongoing challenge of mapping natural language to symbolic logic.

Overall, these results suggest that representation choice matters more than model size. While larger proprietary models perform strongly in zero-shot settings, they still exhibit bias (Tables 1, 2). In contrast, smaller models guided by explicit structural supervision can match or exceed them in logical robustness (Tables 1, 2). This underscores the importance of symbolic structure for bias-resilient reasoning, especially in multilingual settings.

6 Conclusion

In this work, we study content bias in logical reasoning through SemEval-2026 Task 11 and show that explicitly modeling logical structure enables bias-resilient inference. While direct validity prediction remains sensitive to semantic content, our structure-first approach abstracts syllogisms into symbolic representations, achieving strong and consistent performance across both English and multilingual settings. We further demonstrate that structural reasoning signals can be distilled into smaller open-weight models, enabling robust logical performance without reliance on large proprietary systems. Overall, our findings highlight the importance of representation design over model scale and suggest a simple, effective path toward more reliable and bias-resistant reasoning in LLMs.

7 Limitations

Our approach assumes standard three-statement syllogisms with two premises and a conclusion. Extending it to more complex arguments with multiple premises or sorites would require more advanced structural extraction methods.

In the multilingual setting, our distillation experiments are limited to Qwen-3 14B. It remains unclear whether the same structural supervision transfers equally well to other open-weight models, and further evaluation across architectures is needed.

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 Pro-*

cessing, pages 13882–13905, Miami, Florida, USA. Association for Computational Linguistics.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

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.

Philipp Mondorf and Barbara Plank. 2024. Comparing inferential strategies of humans and large language models in deductive reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9402, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2025. SylloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7235–7258, Albuquerque, New Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yinsheng Zhang and Xiaodong Qiao. 2009. A formal system of aristotelian syllogism based on automata grammar. volume 5, pages 173–178.

A Appendix

This appendix details all prompt templates used across our experimental pipeline, organized by task and modeling strategy.

A.1 Subtask 1: Two-Step Structure-First Pipeline (Gemini-3 Pro Preview)

This two-step pipeline using Gemini-3 Pro Preview achieved a perfect combined score of 100 on the Subtask 1 private test set.

A.1.1 Step 1: Cleaned Syllogisms Extraction

This prompt is used to convert raw conversational syllogisms into standardized three-sentence categorical form before logical analysis.

Syllogism Normalization Prompt

```
You are a precise Text Normalizer for logical arguments.

### INPUT:
{syllogism}

### YOUR TASK:
Rewrite the input text into exactly three clean, standard sentences.

### RULES:
1. **Structure:**
   - Sentence 1: Premise 1
   - Sentence 2: Premise 2
   - Sentence 3: Conclusion
2. **Remove "Fluff":** Delete introductory phrases like "It is true that," "We can
   see," "It is a fact," or "Without exception" (unless it changes the quantifier).
3. **Preserve Logic:**
   - Keep Quantifiers EXACTLY as they are (All, No, Some, Many, A few).
   - Keep Negations EXACTLY as they are (not, non-, never).
   - Do NOT change nouns (e.g., do not change "automobiles" to "cars").
4. **Standardize Conclusion:**
   - The third sentence MUST begin with "Therefore,".

### OUTPUT FORMAT:
Return ONLY the three sentences separated by a newline. Do not add labels like
"P1:" or "Conclusion:".
```

A.1.2 Step 2: Symbolic Structural Extraction

This prompt is used to parse the cleaned syllogisms produced in Step 1 and extract their Aristotelian mood-figure codes for formal validity verification.

Symbolic Structure Extraction Prompt

```
You are an expert Formal Logic Parser specialized in Aristotelian Syllogisms.
Your task is to analyze a "Cleaned Syllogism" and convert it into its Abstract Code
(e.g., "AAA-1", "EAE-2").

### PART 1: DEFINITIONS (CONTEXT)
**A. THE MOOD (Sentence Types)**
- **Type A (Universal Affirmative):** All, Every, Each, Any.
- **Type E (Universal Negative):** No, None, Not a single.
- **Type I (Particular Affirmative):** Some, A few, Many, Several, At least one.
- **Type O (Particular Negative):** Some... are not, Not all.

**B. THE FIGURE (Position of M)**
*Let Sub = Subject position, Pred = Predicate position.*
- **Figure 1:** M is Sub in P1 AND Pred in P2. (Structure: M-P, S-M)
- **Figure 2:** M is Pred in P1 AND Pred in P2. (Structure: P-M, S-M)
- **Figure 3:** M is Sub in P1 AND Sub in P2. (Structure: M-P, M-S)
- **Figure 4:** M is Pred in P1 AND Sub in P2. (Structure: P-M, M-S)

### PART 2: INSTRUCTIONS
1. Look at the LAST sentence (Conclusion) to identify S and P.
2. Look at the first two sentences to find M (the repeated term).
3. Determine the Mood (A, E, I, O) for P1, P2, and Conclusion.
4. Determine the Figure (1, 2, 3, 4) based on M's position.

### INPUT SYLLOGISM:
{cleaned_syllogism}

### OUTPUT FORMAT (Strict JSON):
Return JSON ONLY. No markdown.
{
  "terms": { "S": "text", "P": "text", "M": "text" },
  "abstract_code": "MOOD-FIGURE"
}
Example Code: "EAE-1"
```

A.2 Subtask 1: Training Data Generation Pipeline

This prompt is used to independently classify each individual statement in a syllogism into one of the four Aristotelian categorical types (A, E, I, or O), strictly by logical meaning rather than surface wording, ensuring accurate mood labels for constructing training targets.

Sentence-Level Mood Classification Prompt

```
You are an expert Formal Logic Classifier.

Your task is to determine the Aristotelian categorical mood (A, E, I, or O) of a
single sentence. Classify strictly by logical meaning, not by wording or real-world
truth.

DEFINITIONS

- A (Universal Affirmative):
  Every member of the subject class is included in the predicate class.

- E (Universal Negative):
  Every member of the subject class is excluded from the predicate class.

- I (Particular Affirmative):
  At least one member of the subject class exists and is included in the
  predicate class.

- O (Particular Negative):
  At least one member of the subject class exists and is excluded from the
  predicate class.
  Statements of the form "Not all S are P" mean "Some S are not P".

DECISION PROCEDURE

1. Does the sentence assert existence of at least one subject?
   - If yes, the mood must be I or O.
2. Does the sentence make a claim about all members of the subject?
   - If yes, the mood must be A or E.
3. Is the relationship inclusion or exclusion?
   - If the relationship described is one of inclusion, the statement is
     affirmative (A or I).
   - If the relationship described is one of exclusion, the statement is
     negative (E or O).
4. Universal exclusion of the entire class is E.
   Particular exception to inclusion is O.

IMPORTANT CLARIFICATIONS

- "Not all", "not every", and "not always" are O, not E.
- "Each A is not B" means no A is B (E).
- Ignore rhetorical framing and factual truth.
- Paraphrases must be reduced to standard categorical form before classification.

INPUT SENTENCE:
{sentence}
```

A.3 Subtask 1: Structure-First Fine-Tuning Prompt for SLMs

This prompt is used to train SLMs to perform the complete structure-first pipeline in a single forward pass.

Structure-First Fine-Tuning Prompt

```
You are a Formal Syllogism Normalization, Extraction, and Classification Engine.

Your task is to take a raw syllogism consisting of three sentences and produce:
1. A fully normalized syllogism in canonical Aristotelian categorical form.
2. The extracted logical terms S, P, and M derived strictly from the normalized
   syllogism.
3. The syllogism code (Mood-Figure) derived strictly from the normalized
   syllogism.

Follow the steps and definitions below exactly and in order.
---
INPUT
---
Raw Syllogism:
{syllogism}
---
CATEGORICAL PROPOSITION TYPES (A / E / I / O)
---
A — Universal Affirmative (All S are P)
E — Universal Negative (No S are P)
I — Particular Affirmative (Some S are P)
O — Particular Negative (Some S are not P)
---
MANDATORY EXECUTION ORDER
---
STEP 1: IDENTIFY THE CONCLUSION AND MIDDLE TERM (M)
- The conclusion is the sentence starting with "Therefore", "So", "Hence", or is
  logically the final inference.
- Identify S (Subject of Conclusion) and P (Predicate of Conclusion).
- Examine the other two sentences (Premises).
- M is the concept appearing in BOTH premises but NOT in the conclusion.

STEP 2: UNIFY TERM STRINGS (CRITICAL)
- Select ONE canonical string for S, ONE for P, and ONE for M.
- **Normalization Rule:** Convert singular terms to plural if grammatically
  natural (e.g., "cat" -> "cats").
- Example: If text has "dog" and "dogs", you must convert BOTH to "dogs".

STEP 3: NORMALIZE SENTENCES
- Rewrite all 3 sentences using ONLY the templates:
  - "All [Term1] are [Term2]."
  - "No [Term1] are [Term2]."
  - "Some [Term1] are [Term2]."
  - "Some [Term1] are not [Term2]."
- Prefix the conclusion with "Therefore,".
- Ensure the Major Premise (containing P) is Sentence 1.
- Ensure the Minor Premise (containing S) is Sentence 2.
- Ensure the Conclusion is Sentence 3.

STEP 4: DETERMINE THE MOOD
- Identify the type (A/E/I/O) of Sentence 1, 2, and 3.

STEP 5: DETERMINE THE FIGURE
Analyze the position of M in the normalized premises:

- Figure 1: M is Subject of Maj, Predicate of Min. (Sub-Pred)
- Figure 2: M is Predicate of Maj, Predicate of Min. (Pred-Pred)
- Figure 3: M is Subject of Maj, Subject of Min. (Sub-Sub)
- Figure 4: M is Predicate of Maj, Subject of Min. (Pred-Sub)

STEP 6: CONSTRUCT CODE
Format: [Mood]-[Figure] (e.g., "AAA-1")

---
OUTPUT FORMAT (STRICT JSON)
---
Return valid JSON only.

{
  "cleaned_syllogism": "Sentence 1. Sentence 2. Therefore, Sentence 3.",
  "terms": {"S": "string","P": "string","M": "string"},
  "syllogism_code": "Mood-Figure"
}
```

A.4 Subtask 3: Multilingual Pipeline (Gemini-3 Pro Preview)

This prompt is used to extract syllogistic terms (S, P, M) and mood-figure codes directly from non-English syllogisms without intermediate translation, following the second variant of our multilingual structure-first approach.

Native-Language Structure Extraction Prompt

```
You are a Formal Logic Engine specialized in Multilingual Syllogistic Analysis.
Your objective is to parse a raw syllogism and extract its formal logical skeleton
(Terms, Mood, and Figure) while keeping the core terms in their original language.

---

### 1. CORE DEFINITIONS
Refer to these definitions for every analysis:
* **S (Minor Term):** The grammatical Subject of the third sentence (Conclusion).
* **P (Major Term):** The grammatical Predicate of the third sentence (Conclusion).
* **M (Middle Term):** The term that appears in both Premise 1 (P1) and Premise 2
  (P2), but is absent from the Conclusion.
* **Moods:**
    * [A]: Universal Affirmative (All, Every, Each)
    * [E]: Universal Negative (No, None, Not a single)
    * [I]: Particular Affirmative (Some, Many, A few, There exists)
    * [O]: Particular Negative (Some are not, Not all, Not every)
* **Figures (Position of M):**
    * Figure 1: M is Subject of P1, Predicate of P2.
    * Figure 2: M is Predicate of P1, Predicate of P2.
    * Figure 3: M is Subject of P1, Subject of P2.
    * Figure 4: M is Predicate of P1, Subject of P2.

---

### 2. EXECUTION STEPS (FOLLOW IN ORDER)
1. **Term Identification:** Locate S, P, and M in their original script. Remove
   quantifiers (like "all" or "some"). Keep the terms in their native language.
2. **Lemma Extraction:** In languages with complex grammar (like Russian or
   Bengali), extract the base root of the word. Ensure the Middle Term (M) is
   identified as the same word in both premises.
3. **Mood Assignment:** Label the first premise, second premise, and conclusion as
   A, E, I, or O based on their logical quantifiers.
4. **Figure Mapping:** Determine the position of M in the first two premises to
   find the Figure (1, 2, 3, or 4). Use the native word order; do not rearrange
   the sentence.
5. **Code Compilation:** Combine the moods and the figure into a single code
   (e.g., "EIO-2").

---
### 4. UNIVERSAL CONSTRAINTS
* **Strict Native Terms:** S, P, and M must be returned in their original script
  (Bengali, Chinese, etc.).
* **Geometric Fidelity:** Identify the Figure based on the original syntax. Do not
  move terms to make them sound like English.

---

### 5. TARGET SYLLOGISM FOR ANALYSIS
{syllogism}
```

A.5 Subtask 3: Multilingual Data Augmentation - Translation Prompt (Portuguese Example)

We create multilingual training corpora by translating English syllogisms into 11 target languages. This prompt is used to translate English syllogisms into target languages while preserving exact logical operators, quantifiers, and sentence structure.

Translation Prompt (Portuguese)

```
You are a precise logic-aware translator.
Your task is to translate an English syllogism into Portuguese while preserving the
exact logical structure.

Strict Constraints:

1) Preserve ALL logical operators and quantifiers exactly. Use the following
   mappings:
    - The English words "All", "Every", and "Each" must be translated as "Todo",
      "Toda", "Todos", "Todas", or "Cada".
    - The English words "Some", "A number of", and "There exists" must be
      translated as "Alguns", "Algumas", "Existe", "Existem", or
      "Um certo número de".
    - The English words "No", "None", and "Nothing" must be translated as
      "Nenhum", "Nenhuma", or "Nada".
    - The expression "Not all" must be translated as "Nem todos" or "Nem todas".

2) Do NOT paraphrase or weaken/strengthen any quantifier, negation, or logical
   relation. That means:
   - Never replace "Some" with "Vários" or "Muitos".
   - Never replace "All" with "A maioria".
   - Never replace "No" with "Quase nenhum" or similar.

3) Preserve the original sentence order:
   Premise 1 → Premise 2 → Conclusion.
   (In Portuguese, keep this same sequence.)

4) For the conclusion marker, use the direct translation of the marker in the
   English syllogism:
    - "Therefore" → "Portanto"
    - "Consequently" / "As a consequence" → "Consequentemente"
    - "Thus" → "Assim"
    - "It follows that" → "Segue-se que" (only if explicitly present in English)

5) Do NOT explain anything.
   Do NOT add commentary, notes, bullet points, or formatting.
   Output ONLY the Portuguese syllogism as natural Portuguese sentences.

6) **Every English noun phrase, verb, and adjective must be translated directly
   and literally.**
   - Do NOT use synonyms.
   - Do NOT introduce new words.
   - Do NOT remove words.
   - Use only the direct standard Portuguese equivalent.
   - Maintain correct gender and number agreement in Portuguese (e.g.,
     "Todas as coisas", "Algumas formas").

Examples for guidance:

English Syllogism: All humans are mortal. Socrates is a human. Therefore, Socrates
is mortal.
Portuguese Syllogism: Todos os humanos são mortais. Sócrates é um humano.
Portanto, Sócrates é mortal.

English Syllogism: No planets are stars. Some celestial bodies are planets. Thus,
some celestial bodies are not stars.
Portuguese Syllogism: Nenhum planeta é uma estrela. Alguns corpos celestes são
planetas. Assim, alguns corpos celestes não são estrelas.

Now translate the following syllogism into Portuguese:

English Syllogism: {syllogism}
```

A.6 Subtask 3: Syllogism Structure Extraction Prompt (Portuguese Example)

This prompt is used to extract mood-figure codes from machine-translated syllogisms while simultaneously producing cleaned, normalized versions with lemmatized terms and noise removal, using language-specific few-shot examples to guide both parsing and normalization.

Syllogism Structure Extraction Prompt (Portuguese)

```
You are a Portuguese Logic & Linguistics Expert. Analyze the provided Portuguese
syllogism and extract its formal logical structure while also producing a cleaned,
normalized version.

### EXECUTION ORDER
1. **Clean and Normalize the Syllogism:**
   - Remove introductory phrases (e.g., "É verdade que", "Podemos ver que").
   - Standardize the conclusion marker to "Portanto,".
   - Lemmatize all terms to their base (singular) form.
   - Ensure exactly three sentences: Premise 1, Premise 2, Conclusion.
   - Preserve all quantifiers exactly (Todos, Nenhum, Alguns, etc.).

2. **Analyze Conclusion:** Identify the FULL conclusion statement (the last
   sentence). Extract S (Subject) and P (Predicate) in their lemmatized form.

3. **Analyze Premises:** Identify the Middle Term (M) shared between the first two
   sentences. Lemmatize M to its base form.

4. **Map the Figure:** Determine the position of M in the cleaned premises:
   - Figure 1: M is Subject of Premise 1, Predicate of Premise 2.
   - Figure 2: M is Predicate of both premises.
   - Figure 3: M is Subject of both premises.
   - Figure 4: M is Predicate of Premise 1, Subject of Premise 2.

5. **Identify Mood:**
   - A (Todos / Todas)
   - E (Nenhum / Nenhuma)
   - I (Alguns / Algumas)
   - O (Alguns ... não)

### FEW-SHOT EXAMPLES

Example 1:
Input: "Todos os seres humanos são mortais. Sócrates é um ser humano. Logo,
Sócrates é mortal."
Result: {
    "cleaned_syllogism": "Todos os seres humanos são mortais. Sócrates é um ser
    humano. Portanto, Sócrates é mortal.",
    "S": "Sócrates",
    "P": "mortal",
    "M": "ser humano",
    "mood": "AAA",
    "figure": "1"
}

Example 2:
Input: "É verdade que todos os pássaros têm penas. Todos os pássaros põem ovos.
Logo, alguns poedores de ovos têm penas."
Result: {
    "cleaned_syllogism": "Todos os pássaros têm penas. Todos os pássaros põem
    ovos. Portanto, alguns poedores de ovos têm penas.",
    "S": "poedor de ovos",
    "P": "pena",
    "M": "pássaro",
    "mood": "AAI",
    "figure": "3"
}

### YOUR TASK
Syllogism: {syllogism}
```

A.7 Subtask 3: Fine-Tuning Prompt for SLM (Multilingual)

This prompt is used to fine-tune Qwen-3 14B for cross-lingual syllogistic analysis, requiring term extraction and code generation across 11 languages in a single model.

Multilingual Structure-First Fine-Tuning Prompt

```
Analyze a non-English syllogism and extract its formal logical skeleton.
Follow these rules strictly. Do not explain your reasoning.

---
INPUT
---
Syllogism: {syllogism}

---
LOGICAL DEFINITIONS
---
- S (Minor Term): Subject of the Conclusion.
- P (Major Term): Predicate of the Conclusion.
- M (Middle Term): Term in both premises, but NOT the conclusion.

MOODS (Identify based on native logical function)
- A: Universal Affirmative (Total Inclusion).
    Action: Identify native words representing "All", "Every", "Each".
- E: Universal Negative (Total Exclusion).
    Action: Identify native words representing "No", "None", "Nobody".
- I: Particular Affirmative (Partial Inclusion).
    Action: Identify native words representing "Some", "A few", "At least one".
- O: Particular Negative (Partial Exclusion).
    Action: Identify native words representing "Some... are not", "Not all".

FIGURES (Based on M position):
- Fig 1: M-Subj Major / M-Pred Minor
- Fig 2: M-Pred Major / M-Pred Minor
- Fig 3: M-Subj Major / M-Subj Minor
- Fig 4: M-Pred Major / M-Subj Minor

---
EXECUTION STEPS
---
1. IDENTIFY CONCLUSION: Locate the inference (via markers like "Therefore"
   or the final sentence). Extract S and P.
2. IDENTIFY PREMISES: Locate Major (contains P) and Minor (contains S).
   Identify M (the shared term).
3. NORMALIZE (CRITICAL): Extract S, P, and M in Base/Lemma form
   (Singular, Nominative). Keep native script; do NOT translate.
4. DETERMINE MOODS: Assign A/E/I/O to Major, Minor, and Conclusion
   based on native grammar/negation.
5. DETERMINE FIGURE: Match M's role in premises to Fig 1-4 definitions.
6. CONSTRUCT CODE: Format as Mood-Figure (e.g., AAA-1).

---
OUTPUT FORMAT (STRICT JSON)
---
Return ONLY valid JSON:
{
  "S": "string", "P": "string", "M": "string", "syllogism_code": "Mood-Figure"
}
```