SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Large Language Models

Marco Valentino[1], Leonardo Ranaldi[2],
Giulia Pucci[3], Federico Ranaldi[4], André Freitas[5,6]
[1]University of Sheffield, UK [2]University of Edinburgh, UK
[3]University of Aberdeen, UK [4]University of Rome Tor Vergata, Italy
[5]Idiap Research Institute, Switzerland [6]University of Manchester, UK

Abstract

SemEval-2026 Task 11 evaluates the ability of Large Language Models (LLMs) to perform content-independent reasoning through a novel multilingual syllogistic dataset designed to measure the "content effect" — the tendency to conflate semantic plausibility with logical validity. The competition featured four subtasks, covering English and multilingual settings with both standard and noisy premise sets. Evaluations of zero-shot baselines reveal that the content effect is pervasive in open models, whereas newer versions demonstrate a significant shift in performance. Across the subtasks, findings indicate that introducing distracting premises can challenge the models' ability to filter misleading information, while multilingual settings amplify their susceptibility to content biases compared to English. Participants proposed diverse approaches, including neuro-symbolic decomposition, fine-tuning and distillation methods, data augmentation, and activation steering. While explicit symbolic verification remains the most reliable strategy, activation-level interventions and fine-tuning methods offer promising pathways for internalising formal logic within neural architectures. Ultimately, the task reinforces the efficacy of neuro-symbolic approaches and emerging architectural trends for logical reliability, while also highlighting that multilingual setups and longer contexts still pose significant challenges to be investigated in future research.[1]

1 Introduction

The extent to which Large Language Models (LLMs) can learn generalisable, content-independent reasoning mechanisms is still largely disputed in the NLP community (Song et al., 2026; Seals and Shalin, 2024; Wysocka et al., 2024; Ozeki et al., 2024; Dasgupta et al., 2022; Bertolazzi et al., 2024; Eisape et al., 2024; Maraia et al., 2026b; Mirzadeh et al., 2025). Recent work, for example, has shown that LLMs suffer from content biases when assessing or formulating logical arguments — they tend to overestimate the formal validity of arguments that are compatible with world knowledge, and underestimate the formal validity of less plausible arguments (Dasgupta et al., 2022; Valentino et al., 2026; Kim et al., 2025; Bertolazzi et al., 2024, 2025; Balappanawar et al., 2025).

This phenomenon, known as the *content effect*, suggests that formal reasoning in LLMs is inherently entangled with the world knowledge acquired during pre-training. Such entanglement contributes to significant limitations, including susceptibility to spurious correlations and systematic biases in decision-making, which ultimately hinder the deployment of LLMs in critical real-world applications. Because of its impact on reliability, several works have proposed solutions for disentangling content from formal reasoning, including neuro-symbolic frameworks (Quan et al., 2024; Pan et al., 2023; Kambhampati et al., 2024), prompting techniques (Lyu et al., 2023; Xu et al., 2024), supervised fine-tuning via symbolic demonstrations (Ranaldi et al., 2025; Tan et al., 2025), and activation steering methods (Valentino et al., 2026; Maraia et al., 2026a,b; Bertolazzi et al., 2025; Fang et al., 2026). Despite these efforts, a universal solution remains elusive. Furthermore, current studies are predominantly in English, leaving multilingual settings underdeveloped.

To address and improve our understanding of this persisting challenge, we proposed *SemEval-2026 Task 11: Disentangling Content and Formal Reasoning in Large Language Models*. This shared task presents a multilingual evaluation framework for syllogistic reasoning where models are required to assess the formal validity of arguments regardless of their semantic plausibility. We introduce

[1]SemEval-2026 Task 11 website: https://sites.google.com/view/semeval-2026-task-11. GitHub repository: https://github.com/neuro-symbolic-ai/semeval_2026_task_11

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3716–3730

Condition	Plausible	Implausible
Valid	**P1:** All mammals are animals. **P2:** All dogs are mammals. **Conclusion:** All dogs are animals.	**P1:** All birds are reptiles. **P2:** All dogs are birds. **Conclusion:** All dogs are reptiles.
Invalid	**P1:** All flowers are plants. **P2:** All roses are plants. **Conclusion:** All roses are flowers.	**P1:** All planets are round. **P2:** All coins are round. **Conclusion:** All coins are planets.

Table 1: Examples of syllogistic arguments intersecting formal validity and semantic plausibility. The four quadrants illustrate the experimental conditions used to measure the content effect: the tendency for models to conflate the logical structure of an argument ($P \models c$) with the real-world truth of its premises and conclusion.

a novel dataset spanning 24 syllogistic schemes across ten diverse languages. The task is structured into four distinct subtasks:

1. *Subtask 1: Syllogistic Reasoning in English*, focuses on predicting the logical validity of syllogisms independently of semantic plausibility[2].

2. *Subtask 2: Syllogistic Reasoning with Irrelevant Premises in English*, challenges model robustness by requiring the identification of necessary premises amidst "noisy" or logically irrelevant statements[3].

3. *Subtask 3: Multilingual Syllogistic Reasoning*, extends the core validity classification task to multiple diverse languages [4].

4. *Subtask 4: Multilingual Syllogistic Reasoning with Irrelevant Premises*, combines the challenges of multilingual transfer and robustness against irrelevant premises [5].

The competition was hosted on Codabench (Xu et al., 2022), attracting approximately 250 participants with over 3000 submissions across the subtasks and 35 system description papers.

A first analysis of open-source models tested by participants reveals a significant disparity between raw classification accuracy and logical robustness. Evaluations of zero-shot baselines show that the content effect remains pervasive in widely used models such as `Llama 3` (Grattafiori et al., 2024) and `Qwen 2.5` (Bai et al., 2023), where high semantic plausibility often overrides logical structure. Notably, increased parameter scale does not consistently yield higher robustness, as larger models frequently proved more susceptible to internalizing pre-training biases. However, newer model generations, including `Qwen 3` (Yang et al., 2025) and `Phi-4` (Abdin et al., 2024), demonstrate a significant performance shift, achieving accuracies above 95 with lower content bias.

Participating systems adopted diverse strategies to address the task which can be categorised into five main paradigms: (i) pseudo-formal neuro-symbolic decomposition, (ii) architectural modifications and fine-tuning, (iii) data augmentation and synthetic generation, (iv) latent-representation steering, and (v) task-specific heuristics.

While explicit symbolic verification emerged as the most effective strategy for bypassing the content effect by delegating inference to deterministic solvers, neural-based interventions like activation steering and targeted fine-tuning offer promising pathways for internalising robust, content-invariant reasoning directly within the model's parameters. However, the shared task also clearly demonstrates that filtering distracting premises and maintaining logical consistency across diverse languages remain substantial hurdles.

Ultimately, the task outcomes reinforce the reliability of neuro-symbolic approaches for logical reasoning , while highlighting that recent architectural refinements are beginning to bridge the gap between plausible and formal reasoning, even as multilingual setups and longer, noisy contexts continue to pose significant challenges for future investigation.

[2]Subtask 1 competition: https://www.codabench.org/competitions/11928/

[3]Subtask 2 competition: https://www.codabench.org/competitions/11929/

[4]Subtask 3 competition: https://www.codabench.org/competitions/11930/

[5]Subtask 4 competition: https://www.codabench.org/competitions/11931/

2 Task Description & Objectives

The primary objective of SemEval-2026 Task 11 is to investigate and improve the capabilities of LLMs in formal reasoning. Syllogistic reasoning, a foundational form of deductive logic, serves as the primary vehicle for this investigation. A syllogism typically consists of premises and a conclusion, where the main task is to determine if the conclusion logically follows from the premises.

A core challenge in LLMs' reasoning is the *content effect*, a phenomenon where a model's judgment of logical validity is biased by the semantic plausibility of the argument's content. By presenting syllogistic arguments that are either aligned (plausible) or misaligned (implausible) with world knowledge, we can measure the model's ability to prioritise logical structure over semantic bias. Table 1 illustrates the four dimensions investigated in the task: valid-plausible, valid-implausible, invalid-plausible, and invalid-implausible.

2.1 Problem Formalization

Formally, the task is defined as a binary classification problem. Given a syllogistic argument $A = \{P, c\}$, where $P = \{p_1, p_2, \ldots, p_n\}$ is a set of premises and c is a conclusion, the model must predict a label $y \in \{V, I\}$ representing the *formal validity* of the argument.

- $y = V$ (Valid): The conclusion c is logically entailed by the premises P such that $P \models c$.
- $y = I$ (Invalid): The conclusion c does not logically follow from the premises P.

The fundamental challenge lies in the disentanglement of the validity label y from the *semantic plausibility* (S) of the argument A, where $S \in \{true, false\}$ based on real-world knowledge. The *content effect* occurs when a model's prediction of y is erroneously influenced by S.

2.2 Subtasks Overview

The competition is structured into four subtasks designed to evaluate model capabilities across different languages and complexity levels.

2.2.1 Subtask 1: Syllogistic Reasoning in English

The goal of this subtask is to determine the formal validity of syllogisms presented in English. The task requires models to identify whether a conclusion follows logically from a set of premises based solely on structural deduction, ignoring the semantic plausibility of the content. Evaluation is based on the following metrics:

Accuracy (ACC) The percentage of correct validity predictions across the entire test set.

Intra-Plausibility Content Effect The average difference in accuracy between valid and invalid arguments within a specific plausibility group, measuring bias toward a specific validity label.

Cross-Plausibility Content Effect The average difference in accuracy between plausible and implausible arguments for a fixed formal validity value, measuring bias toward the semantic truth of the conclusion.

Total Content Effect (TCE) The mean of the intra- and cross-plausibility content effects. A lower TCE indicates that the model relies on logical structure rather than real-world knowledge.

Combined Score (CS) Ranking is determined by the ratio of accuracy to the content effect, rewarding models that are both correct and robust:

$$\text{CS} = \frac{\text{ACC}}{1 + \ln(1 + \text{TCE})} \tag{1}$$

2.2.2 Subtask 2: Syllogistic Reasoning with Irrelevant Premises in English

This subtask evaluates model robustness by introducing irrelevant or "noisy" premises. Models must jointly predict the validity of the syllogism and identify the subset of relevant premises necessary and sufficient to entail the conclusion. Notably, in this framework, only valid syllogisms contain relevant premises. The evaluation employs:

Classification Accuracy (ACC) The percentage of correct validity predictions.

Premise F1 Score (F1) The macro-averaged F1-score measuring the model's ability to correctly identify the specific subset of relevant premises.

Total Content Effect (TCE) A composite score measuring the model's susceptibility to semantic bias across different plausibility conditions.

Combined Score (CS) Ranking is based on the ratio of the average (Avg) between Accuracy and F1 to the penalized content effect:

$$\text{CS} = \frac{\text{Avg}(\text{ACC}, \text{F1})}{1 + \ln(1 + \text{TCE})} \tag{2}$$

2.2.3 Subtask 3: Multilingual Syllogistic Reasoning

This subtask extends the binary classification of syllogistic validity to a multilingual setting, covering diverse language families and scripts. The goal is to assess whether logical reasoning capabilities and content biases remain consistent across different languages. Evaluation metrics include:

Classification Accuracy (Acc) The average classification accuracy across all evaluated languages.

Total Content Effect (TCE) A composite score measuring the model's susceptibility to semantic bias across different plausibility conditions and different languages.

Combined Score (CS) Similarly to Subtask 1, participants are ranked according to the ratio of average accuracy to the total content effect:

$$\text{CS} = \frac{\text{Acc}}{1 + \ln(1 + \text{TCE})} \tag{3}$$

2.2.4 Subtask 4: Multilingual Syllogistic Reasoning with Irrelevant Premises

This subtask combines multilingual syllogistic reasoning with the requirement to filter irrelevant contexts. Models must jointly predict logical validity and identify relevant premises across all target languages. Only valid syllogisms contain a set of necessary and sufficient premises. Performance is measured via:

Classification Accuracy (Acc) The average accuracy for validity prediction across all languages.

Premise F1 Score (F1) The macro-averaged F1-score for identifying relevant premises across all languages.

Total Content Effect (TCE) A composite metric measuring susceptibility to semantic bias and stability across the multilingual corpus.

Combined Score (CS) Similarly to Subtask 2, ranking is determined by the ratio of the multilingual combined performance (the mean of Acc and F1) to the content effect:

$$\text{Score} = \frac{\text{Avg}(\text{Acc}, \text{F1})}{1 + \ln(1 + \text{TCE})} \tag{4}$$

2.3 Languages

To evaluate the consistency of the content effect across diverse linguistic contexts, the competition encompasses eleven languages representing a variety of language families, scripts, and resource levels. Subtasks 1 and 2 focus on English (en). Subtasks 3 and 4 expand the evaluation to a multilingual corpus consisting of high-resource languages, including German (de), Spanish (es), French (fr), Italian (it), Dutch (nl), Portuguese (pt), Russian (ru), and Chinese (zh), alongside mid- and low-resource languages, such as Swahili (sw), Bengali (bn), and Telugu (te).

3 Dataset Construction

English Dataset We constructed a dataset comprising syllogistic arguments following the 24 valid syllogistic schemes under Aristotelian interpretation with existential import. The starting point for constructing the dataset was a collection of syllogistic reasoning problems in English, selected to maintain controlled logical structures across content variations. Inspired by previous work in the field (Bertolazzi et al., 2024; Valentino et al., 2026; Wysocka et al., 2024), we constructed our dataset starting from the symbolic logical representation of the 24 valid schemes and a set of known logical fallacies from which we derive structured natural language templates (Appendix A).

We then leveraged the templates to construct logically valid and invalid arguments (based on the conclusions) and instantiate them with plausible and implausible content using Gemini 3.0 Pro (Team et al., 2023).

```
syllogism: "All apples are edible
fruits. All edible fruits are fruits.
All apples are fruits"
validity: true
plausibility: true
```

Subsequently, we use Gemini again to paraphrase the syllogisms and increase linguistic diversity.

```
syllogism: "It is true that every
apple is an edible fruit. Each edible
fruit, without exception, is a fruit.
Therefore, every apple is a fruit."
validity: true
plausibility: true
```

For Subtasks 2 and 4, along with paraphrasing, we use Gemini to generate a set of logically irrelevant premises whose content is semantically similar

Model	Acc ↑	TCE ↓	CS ↑
Qwen 3 14B	97.38	4.23	36.62
Qwen 3 8B	95.81	4.23	36.09
Phi 4 15B	83.33	14.89	22.13
Qwen3 8B	78.00	21.10	19.00
Qwen 2.5 7B Instruct	74.31	31.20	16.62
Gemma3 7B	71.70	29.75	16.20
Qwen 2.5 3B Instruct	65.28	19.86	16.17
Qwen2.5 7B Instruct	67.02	32.69	14.84
Phi-4 4B Mini Instruct	70.83	44.61	14.70
Llama 3.1 8B	64.17	32.15	14.26
Qwen2.5 14B	67.50	44.54	14.01
SmolLM3 3B	56.54	24.51	13.34
Llama 3.2 3B Instruct	57.64	29.81	13.02
Llama 3.1 8B Instruct	58.33	49.15	11.87

Table 2: Reasoning accuracy and content biases of base language models identified by participants on Subtask 1 without any specific architectural interventions, ordered by final Combined Score (CS).

to the core syllogistic argument. Finally, we performed a manual validation to maximise quality and correct errors introduced during the automatic generation process.

The final dataset consists of 960 training examples and 192 test examples in English for Subtask 1 and 192 for Subtask 2, equally distributed between valid-plausible, valid-implausible, invalid-plausible, and invalid-implausible arguments.

Multilingual Dataset To evaluate multilingual transfer, we expanded the dataset into nine additional languages spanning diverse families, scripts, and resource levels (Section 2.3). The multilingual dataset is generated starting from English examples, translating them via GPT-4o (Hurst et al., 2024). To ensure that the logical structures were preserved, we implemented a rigorous three-stage quality-control pipeline. This process included structural guards, an automatic back-translation validation step evaluated via SBERT cosine similarity (Reimers and Gurevych, 2019), and a final bilingual human review to verify the faithful rendering of quantifiers and the strict isomorphism of logical forms. A total of 192 test examples were generated for each Subtask.

4 Results

The competition was hosted on Codabench (Xu et al., 2022), attracting approximately 250 participants with over 3000 submissions across the subtasks and 35 system description papers. Here, we highlight the main findings and trends emerging from the shared task.

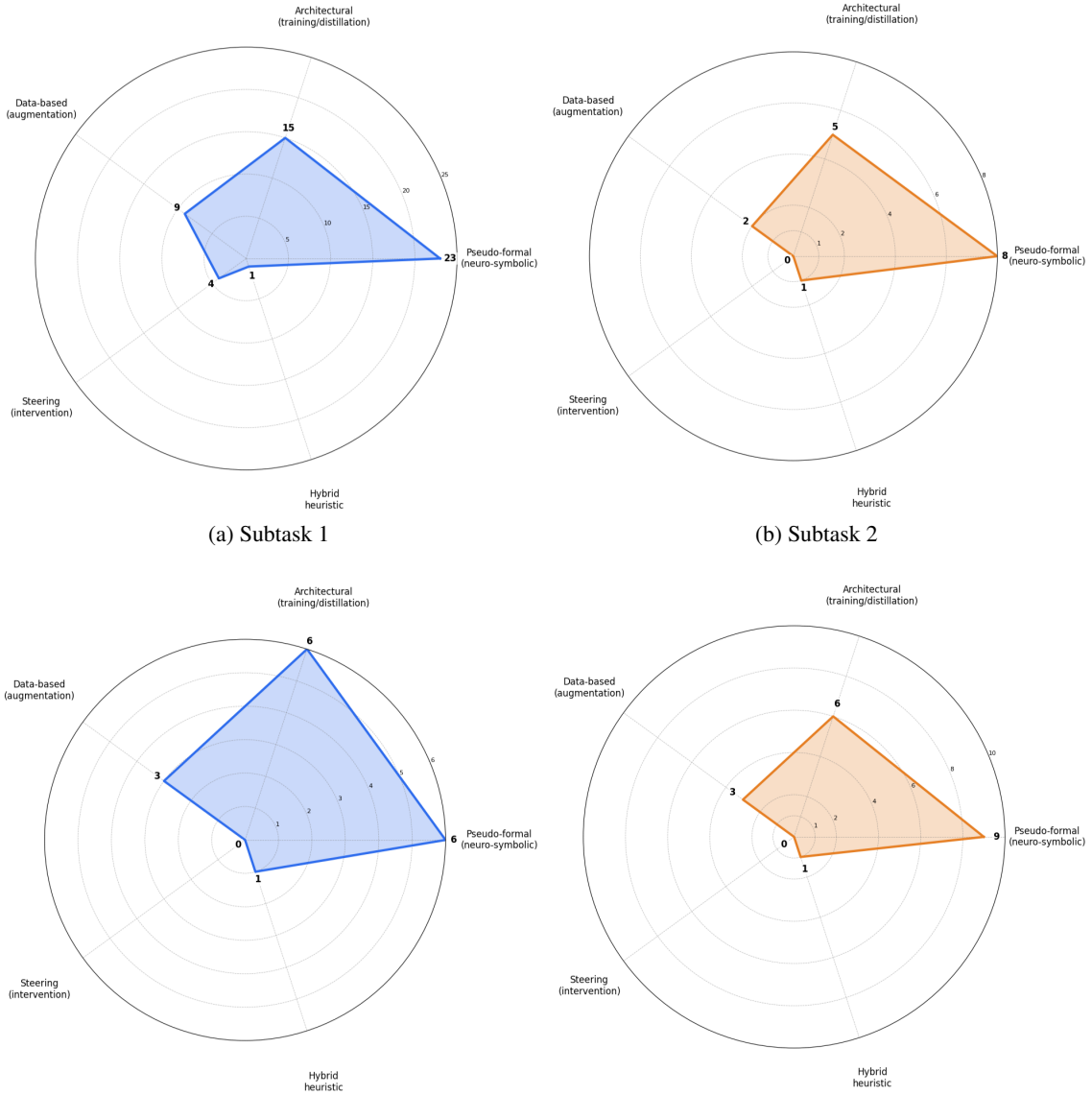
4.1 Content Effect in Language Models

To evaluate the baseline logical integrity of current open-source architectures, we analyse the content effects exhibited by various model families and sizes as reported by participants using zero-shot or standard prompting without specialised interventions. The results, summarised in Table 1, reveal a significant disparity between raw classification accuracy and the ability to maintain structural reasoning under semantic pressure.

Across the majority of mid-sized models, such as the Llama-3.1 and Qwen-2.5 series, we observe a pervasive entanglement of logic and belief. For instance, while Llama-3.1-8B achieves a respectable accuracy of 64.17 , its high Total Content Effect (TCE) of 32.15 leads to a significantly lower final score of 14.26. This trend is even more pronounced in the Qwen-2.5-14B base model, where a higher accuracy of 67.50 is undermined by a TCE of 44.54, resulting in a lower overall score than its 8B Llama counterpart. These results suggest that increasing parameter count within the same architectural generation does not necessarily yield a proportional increase in logical robustness.

However, the data also highlights a notable generational shift in reasoning capabilities with the emergence of the Qwen-3 and Phi-4 families. The Qwen-3 14B and 8B models demonstrate a step-change in performance, achieving accuracies above 95 with remarkably low TCE values (4.23). This leads to primary scores (36.62 and 36.09, respectively) that are more than double those of previous-generation models. Similarly, Phi-4 15B exhibits a strong balance between performance and bias resistance, with an accuracy of 83.33 and a TCE of 14.89. These outliers suggest that recent advancements in pre-training data curation or architectural refinements may be successfully prioritizing structural invariance over semantic pattern matching.

Finally, the performance of smaller models, such as SmolLM3-3B (ACC: 56.54, TCE: 24.51) and Llama 3.2 3B Instruct (ACC: 57.64, TCE: 29.81), indicates that while small-scale models struggle with the complexity of the syllogistic task, they do not necessarily exhibit higher content bias than their 7B or 14B predecessors. Notably, Qwen 2.5 3B Instruct achieves a score of 16.17, nearly matching the 7B Gemma3, primarily due to its relatively low TCE of 19.86. Collectively, these findings confirm that while content-independent reasoning remains a challenge, newer architectural iterations



(a) Subtask 1

(b) Subtask 2

(c) Subtask 3

(d) Subtask 4

Figure 1: Distribution of methodological paradigms adopted by participating systems for Subtasks 1 and 2. The radar chart shows the number of papers employing each type of approach.

are beginning to bridge the gap between semantic understanding and logical deduction.

4.2 Methodological Paradigms

The submissions to the shared task adopted a wide range of strategies for disentangling formal logical validity from semantic plausibility. Despite their methodological diversity, most systems can be grouped into a small number of paradigms according to how they attempt to mitigate the *content effect*. Broadly, we identify five families of approaches: (i) pseudo-formal approaches based on neuro-symbolic task decomposition, (ii) architectural approaches that modify the model or its training process, (iii) data-based approaches relying on augmentation and synthetic data generation, (iv) latent-representation interventions such as activation steering, and (v) other heuristic strategies.

The distribution of architectural trends and strategies adopted by the participants across the different subtasks is illustrated in Figure 1.

Pseudo-formal approaches: neuro-symbolic task decomposition and theorem provers. A large portion of submissions adopts a pseudo-formal paradigm in which the reasoning process is decomposed into two stages: natural language interpretation followed by formal logical verification. In these systems, a language model is typically

ID	Authors	Approach	Acc	TCE	CS
Subtask 1					
46	(Luo et al., 2026)	P+D	100	0	100
60	(Fu et al., 2026)	P+A	100	0	100
82	(Li et al., 2026)	D	100	0	100
89	(Advani, 2026)	P	100	0	100
104	(Tran et al., 2026)	A+D	100	0	100
141	(Wang et al., 2026)	P+A	100	0	100
156	(Shaikh et al., 2026)	P+A	100	0	100
248	(Muhamad et al., 2026)	P	100	0	100
256	(R and Chopra, 2026)	P+A	100	0	100
466	(Butas et al., 2026)	P+A+D	100	0	100
305	(Księżniak, 2026)	P+A+D	97,91	0,02	95,81
184	(Huiskens et al., 2026)	S	98,75	1,58	50,60
18	(Sharma et al., 2026a)	P	97,38	3,10	50,24
10	(Setu and Malhotra, 2026)	A+D	94,27	3,21	48,74
208	(Grimaldi, 2026)	P	96,34	1,02	46,57
204	(Ayman et al., 2026)	P	98,95	2,13	46,43
441	(Petersen et al., 2026)	P	95,29	2,15	44,47
469	(Krishnasamy, 2026)	P+A	94,24	2,08	44,30
160	(Tratzsch et al., 2026)	S	82,30	2,08	42,19
365	(Akinfaderin and Diallo, 2026)	P	94,30	2,85	41,88
431	(Gogu et al., 2026)	P+A	98,43	3,19	40,45
72	(de Rijke et al., 2026)	P+A	96,34	3,12	39,86
206	(Chen et al., 2026)	A+D	96,86	3,19	39,81
461	(Kartáč et al., 2026)	P	95,29	3,21	39,08
79	(Faisal and Chowdhury, 2026)	P	93,19	3,12	38,56
303	(Memar et al., 2026)	A+D	82,72	3,12	34,22
351	(Chowdhury et al., 2026)	P+A+S	86,81	4,60	31,16
353	(Sunil Saumya and Sai Reddy, 2026)	A+D	79,06	4,17	29,92
274	(Gupta et al., 2026a)	P+H	91,58	7,33	28,08
319	(Abi Akl et al., 2026)	P+A	90,58	8,55	27,81
344	(López-Ponce et al., 2026)	P+A+D	72,25	11,77	20,37
309	(Thiyagarajaa and Thenmozhi, 2026)	P	71,73	11,84	20,19
31	(Takahashi, 2026)	D	72,80	31,60	16,20
Subtask 3					
156	(Shaikh et al., 2026)	P+A	100	0	100
248	(Muhamad et al., 2026)	P	100	0	100
256	(R and Chopra, 2026)	P+A	100	0	100
60	(Fu et al., 2026)	P+A	95,30	0,17	92,59
46	(Luo et al., 2026)	P+D	96,35	1.0	56,97
466	(Butas et al., 2026)	P+A+D	98,44	1,09	56,71
19	(Sharma et al., 2026b)	H	90,62	10,42	50,52
251	(Gupta et al., 2026b)	P	95,83	5,21	33,91
141	(Wang et al., 2026)	P+A	94,79	6,25	31,80
461	(Kartáč et al., 2026)	P	93,75	6,25	31,45
441	(Petersen et al., 2026)	P+H	88,02	6,38	29,36
206	(Chen et al., 2026)	A+D	91,67	11,46	26,02

Table 3: Performance Comparison for Subtask 1 (English) and 3 (Multilingual) ordered by Combined Score (CS). For Subtask 1 we only report the top 15 approaches in the ranking. The full results available in the Appendix. **P**: Pseudo-formal approaches (neuro-symbolic task decomposition, theorem provers); **A**: Architectural approaches (training, fine-tuning, distillation); **D**: Data-based approaches (augmentation, synthetic generation); **S**: Latent representation steering; **H**: Other heuristic approaches.

used to extract or normalise the logical structure of a syllogism (e.g., quantifier type, subject and predicate terms, or first-order logic representations). The resulting representation is then passed to a deterministic symbolic procedure that evaluates logical validity. Typical implementations include mapping syllogisms to Aristotelian A/E/I/O forms and verifying membership in the 24 valid moods and figures, translating statements into first-order logic and applying automated theorem provers, or checking satisfiability over Venn-style set representations. Some systems further extend this paradigm to multilingual settings through translation modules or structural consistency checks across multiple representations.

Architectural approaches: training, fine-tuning, and distillation. Another group of submissions focuses on modifying the architecture or the learning procedure of neural models in order to encourage structure-based reasoning. These approaches include supervised fine-tuning of transformer models on syllogistic datasets, parameter-efficient training methods such as LoRA (Hu et al., 2022) or QLoRA (Dettmers et al., 2023), and knowledge distillation from stronger teacher models that generate structured reasoning traces. Several systems also explore ensemble or role-based architectures in which multiple model instances perform complementary reasoning roles (e.g., a "believer" and a "skeptic") and combine their predictions through

	ID	Authors	Approach	Acc	F1	TCE	CS
Subtask 2							
	141	(Wang et al., 2026)	P+A	100	100	0	100
	156	(Shaikh et al., 2026)	P+A	100	100	0	100
	60	(Fu et al., 2026)	P+A	96,84	94,21	1,13	54,43
	466	(Butas et al., 2026)	P+A+D	96,88	95,83	1,17	54,24
	248	(Muhamad et al., 2026)	P	98,94	95,43	2	46,31
	89	(Advani, 2026)	P	95,26	99,47	2,94	41,08
	461	(Kartáč et al., 2026)	P	97,37	96,84	3,30	39,49
	441	(Petersen et al., 2026)	P	95,79	92,37	3,26	38,43
	18	(Sharma et al., 2026a)	P	82,11	99,47	9,89	26,80
	469	(Krishnasamy, 2026)	P+A	85,79	86,32	11,16	24,60
	431	(Gogu et al., 2026)	P+A	97,37	9,47	3,12	22,10
Subtask 4							
	141	(Wang et al., 2026)	P+A	90,10	89,58	1,26	49,51
	461	(Kartáč et al., 2026).	P	84,90	83,42	1,37	45,20
	248	(Muhamad et al., 2026)	P	90,63	90,10	3	37,88
	156	(Shaikh et al., 2026)	P+A	89,06	89,06	2,89	37,38
	466	(Butas et al., 2026)	P+A	91,15	90,10	5,34	31,83
	60	(Fu et al., 2026)	P+A	85,94	87,50	5,43	30,31
	441	(Petersen et al., 2026)	P	77,08	64,15	6,07	23,89
	19	(Sharma et al., 2026b)	H	74,48	77,60	9,55	22,66

Table 4: Performance Comparison for Subtask 2 (English) and 4 (Multilingual) ordered by Combined Score (CS). **P**: Pseudo-formal approaches (neuro-symbolic task decomposition, theorem provers); **A**: Architectural approaches (training, fine-tuning, distillation); **D**: Data-based approaches (augmentation, synthetic generation); **S**: Latent representation steering; **H**: Other heuristic approaches.

voting or confidence aggregation mechanisms.

Data-based approaches: augmentation and synthetic generation. A complementary line of work attempts to reduce semantic shortcuts by modifying the training data rather than the reasoning mechanism itself. These approaches introduce various forms of data augmentation, including replacing lexical entities with symbolic placeholders, generating pseudo-words, creating counterfactual syllogisms, or synthesising additional examples that preserve logical structure while altering semantic content. By exposing models to structurally equivalent but semantically diverse examples, these systems aim to discourage reliance on real-world plausibility cues and instead promote structure-based generalisation.

Latent representation interventions through steering. A smaller set of submissions attempts to mitigate content bias by intervening directly in the internal representations of language models. Rather than retraining the model or introducing symbolic reasoning modules, these methods apply inference-time techniques such as activation steering or kNN-based conditioning (Valentino et al., 2026). By modifying hidden activations or conditioning predictions on structurally similar examples, these approaches guide the model toward reasoning patterns that prioritize logical structure over semantic plausibility while preserving the underlying model parameters.

Other heuristic approaches. Finally, a few systems employ heuristic reasoning strategies that do not fit neatly into the previous categories. These approaches often combine rule-based components, adaptive inference procedures (translating to English as pivot), or task-specific heuristics designed to improve robustness without relying on symbolic reasoning or extensive retraining.

4.3 Results Across Subtasks

4.3.1 Impact of Distracting Premises

Subtask 1 evaluates the formal validity of syllogisms presented in English. In this setting, several systems achieve perfect performance, reaching 100% accuracy with zero Content Effect (TCE), which results in the maximum Combined Score (CS) of 100. These top-performing systems are predominantly based on pseudo-formal approaches or combinations of pseudo-formal and architectural methods, often complemented with data-based strategies. This indicates that structured reasoning pipelines and explicit decomposition are highly effective when the task is restricted to a single language and controlled conditions.

Subtask 2 evaluates model robustness by introducing irrelevant or “noisy" premises. Models must jointly predict the validity of the syllogism and identify the subset of relevant premises necessary and sufficient to entail the conclusion. Table 6 shows that in Subtask 2, two systems still achieve perfect performance, reaching 100% accuracy and F1 with 0% Total Content Effect (TCE). As in Subtask 1, the best results are obtained by pseudo-formal and architectural approaches, indicating that structured reasoning pipelines and model adaptations are highly effective when the task requires consistent inference across adversarial configurations. However, systems without these structural interventions exhibit increased susceptibility to irrelevant premises, highlighting that filtering necessary premises presents a notable challenge.

Subtask 3 extends the binary classification of syllogistic validity to a multilingual setting, covering diverse language families and scripts. While a few systems still maintain perfect performance, most approaches exhibit a noticeable degradation compared to the English baseline.

In Subtask 4, performance drops across all metrics. While many systems still maintain relatively high accuracy and F1 scores (often between 85% and 91%), the presence of multilingual distracting information leads to a higher TCE, which significantly reduces the Combined Score. For instance, systems that achieve around 90% accuracy may still exhibit TCE values between 3 and 5, reflecting increased susceptibility to irrelevant premises. This suggests that the main challenge in Subtask 4 is not purely logical reasoning but the ability to filter and ignore misleading information when it is presented in languages that are different from English.

In general, approaches combining pseudo-formal reasoning with architectural adaptations remain the most robust overall, whereas heuristic strategies show greater sensitivity to content-based interference.

4.3.2 English vs. Multilingual Performance

Comparing the strictly English environments (Subtasks 1 and 2) with the multilingual environments (Subtasks 3 and 4) highlights clear performance disparities. Multilingual inputs significantly increase the difficulty of the task.

In Subtask 3, accuracy remains relatively high for several systems, often above 90%, but this improvement is frequently accompanied by a higher TCE, indicating a stronger influence of content expressed in different languages. For example, systems achieving over 90% accuracy may still exhibit TCE values between 5 and 11, substantially lowering their Combined Score. Overall, the comparison suggests that the multilingual setting can substantially affect robustness to content biases.

This confirm that controlling for content effect becomes increasingly important as tasks expand to multilingual and more heterogeneous settings.

5 Related Work

Recent research has demonstrated that Large Language Models (LLMs) exhibit *content effects* in formal reasoning tasks, mirroring cognitive biases that may align or differ from those observed in humans (Dasgupta et al., 2022; Valentino et al., 2026; Kim et al., 2025; Mondorf and Plank, 2024; Seals and Shalin, 2024; Eisape et al., 2024). These effects arise when the semantic plausibility of a problem influences the model’s reasoning process, often leading to correct conclusions for plausible statements and systematic errors for implausible but logically valid ones (Bertolazzi et al., 2024).

The foundational work by (Dasgupta et al., 2022) showed that LLMs perform better on reasoning tasks when the content aligns with world knowledge. In their experiments, models were significantly more accurate on syllogistic tasks when the conclusions were semantically plausible, even when this plausibility conflicted with the actual logical validity of the argument. This indicates a bias toward material reasoning — reasoning grounded in semantic associations — rather than formal reasoning based strictly on logical rules. Further work (Bertolazzi et al., 2024, 2025; Valentino et al., 2026; Maraia et al., 2026b; Kim et al., 2025; Balappanawar et al., 2025) systematically evaluated LLMs on a broad suite of tasks and found that performance dropped sharply for arguments that contradicted commonsense knowledge. This reliance on content plausibility suggests that LLMs are susceptible to semantic interference, failing to uphold the norms of formal logic when they conflict with prior knowledge or beliefs.

Despite these advancements, the current literature remains significantly limited in its linguistic scope (Ando et al., 2023; Maraia et al., 2026a). Existing studies have focused almost exclusively on English, leaving a critical gap in our understanding of whether these biases are universal or vary across different linguistic and cultural contexts.

6 Conclusion

The results of SemEval-2026 Task 11 provide a comprehensive diagnosis of formal reasoning in Large Language Models (LLMs).

The tasks revealed that open language models can be biased by a significant content effect, with architectures such as Llama-3.1 and Qwen-2.5 demonstrating a systematic tendency to conflate semantic plausibility with logical validity. While newer model generations like Qwen-3 and Phi-4 exhibit a notable shift toward structural logic, the entanglement of world knowledge and formal deduction remains a fundamental characteristic of current pre-training objectives across languages.

Methodologically, the competition established that the content effect can be alleviated through diverse strategies, most notably via neuro-symbolic pipelines that delegate inference to deterministic solvers. Simultaneously, advancements in latent-representation steering, data augmentation, and fine-tuning demonstrate that models can be effectively nudged toward prioritising structural features without external symbolic components. However, the pronounced performance drop observed in Subtasks 2 and 4 when faced with "logical noise" indicates that identifying and filtering necessary premises amidst distracting information remains a challenge. Furthermore, the contrast in performance between the English and multilingual subtasks highlights that cross-lingual transfer substantially amplifies a model's susceptibility to these semantic biases.

While the current results demonstrate that content-independent reasoning is achievable in modern LLMs, this work serves primarily as a starting point for future research. Evaluating the content effect in more challenging reasoning scenarios, particularly those involving multilingual setups, increased premise counts, and greater contextual complexity, will be essential to developing models that are truly robust and unbiased in real-world and high-stakes environments.

Acknowledgments

We acknowledge IT Services at the University of Sheffield for the provision of the Stanage HPC cluster, which supported the development of the pilot studies that made this shared task possible.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Hanna Abi Akl, Fabien Gandon, Catherine Faron, and Pierre Monnin. 2026. Sef-clgc at semeval-2026 task 11 subtask 1: Logical notation impact on language model performance. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Laksh Advani. 2026. lakshadvani at semeval-2026 task 11: A neuro-symbolic approach to content-independent syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Adewale Akinfaderin and Nafi Diallo. 2026. Fregelogic at semeval 2026 task 11: A hybrid neuro-symbolic architecture for content-robust syllogistic validity prediction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2023. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 1–11, Nancy, France. Association for Computational Linguistics.

Mohamed Ayman, Khaled Marzouk, Abdallah Mashaly, and Ahmed Hereiz. 2026. Proofbusters at SemEval-2026 Task 11: Symbolic Abstraction Approaches for Improving Logical Reasoning in Language Models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ishwar B Balappanawar, Vamshi Krishna Bonagiri, Anish R Joishy, Manas Gaur, Krishnaprasad Thirunarayan, and Ponnurangam Kumaraguru. 2025. If pigs could fly... can llms logically reason through counterfactuals? *arXiv preprint arXiv:2505.22318*.

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

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. 2025. How language models conflate logical validity with plausibility: A representational analysis of content effects. *arXiv preprint arXiv:2510.06700*.

Rafael Butas, Rodica Potolea, Camelia Lemnaru, and Alex Lapusan. 2026. TUCNLP at SemEval-2026 task 11: Neuro-symbolic content stripping for debiased syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Zhanyu Chen, María Teresa Muñoz Martín, Sem Huisman, and Jingjing Lan. 2026. Sylloscope at SemEval-2026 task 11: Decoupling logic from belief via DeepSeek-enhanced distillation in Qwen models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Akash Chowdhury, Vlad Pavlovich Julius Dunfoy, and Sophia Yang Abhiram Borra. 2026. Abstractreasoner at semeval-2026 task 11: Reducing content effects via knowledge distillation and structured reasoning prompts. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

Janiek de Rijke, Niek Biesterbos, and Mark den Ouden. 2026. RvH-40 at SemEval-2026 task 11: Disentangling reasoning from belief through symbolic abstraction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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 8418–8437.

Adnan Faisal and Shiti Chowdhury. 2026. CUET_Luminaries at SemEval-2026 task 11: Disentangling logical validity from semantic plausibility through canonical abstraction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Feihao Fang, My T Thai, and Yuanyuan Lei. 2026. Discovering a shared logical subspace: Steering llm logical reasoning via alignment of natural-language and symbolic views. *arXiv preprint arXiv:2604.19716*.

Junhao Fu, Yun He, Lina Zhao, and Weijuan Li. 2026. YNJTC at SemEval-2026 task 11: A neuro-symbolic hybrid pipeline for content-independent syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Răzvan-Costinel Gogu, Ştefan Plăcintescu, and Sofia-Maria Vultur. 2026. Ghegheghe at semeval-2026 task 11: Decoupling logic from belief with bias-targeted fine-tuning and neuro-symbolic syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pasquale Grimaldi. 2026. pamaldi at semeval-2026 task 11: Neuro-symbolic syllogistic reasoning via llm-guided structure extraction and deterministic validation. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ishita Gupta, Dhruv Goyal, and Jatin Bedi. 2026a. Single-call joint abstraction for robust neuro-symbolic retrieval: 0704mis at semeval-2026 task 11 subtasks 2 & 4. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ishita Gupta, Dhruv Goyal, and Jatin Bedi. 2026b. Team 0704mis at SemEval-2026 task 11: Dual-view consistency testing for content-invariant multilingual syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Twan Huiskens, Tian Niezing, and Koen Snelten. 2026. d'olle grieze at semeval-2026 task 11: Comparing the impact of supervised fine-tuning and activation steering on mitigating content effect bias in syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*.

Ivan Kartáč, Kristýna Onderková, Jan Bronec, Zdeněk Kasner, Mateusz Lango, and Ondřej Dušek. 2026. UFAL-CUNI at SemEval-2026 task 11: An efficient modular neuro-symbolic method for syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.

Saran Krishnasamy. 2026. GigitAI at SemEval-2026 task 11: Hybrid symbolic-neural approach for syllogistic validity classification. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ewelina Księżniak. 2026. Team ewelinaksiez at SemEvAl-2026 task 11: Reducing content bias in syllogistic reasoning via semantic abstraction. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Songhuan Li, Liang Yang, Shengdi Yin, Qiang Zhang, and Hongfei Lin. 2026. dutir_shlee at SemEval-2026 task 11: Symbolic augmentation for content-bias-resistant syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Francisco F. López-Ponce, Lucía Pitarch, Iván Saavedra-Martínez, Ignacio Huitzil, Sergio-Luis Ojeda-Trueba, Fernando Bobillo, and Gemma Bel-Enguix. 2026. GIL-zaragoza at SemEval 2026 task 11: Comparing classification, autoformalization, and ontologies for formal reasoning capabilities. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Rongchuan Luo, Jin Wang, and Xuejie Zhang. 2026. YNU-HPCC at SemEval-2026 task 11: Mitigating content effects in syllogistic reasoning with Qwen2-1.5B-instruct and XLM-ROBERTa-large for English and multilingual tasks. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. 2026a. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2739–2753, Rabat, Morocco. Association for Computational Linguistics.

Gabriele Maraia, Marco Valentino, Fabio Massimo Zanzotto, and Leonardo Ranaldi. 2026b. Abstract activation spaces for content-invariant reasoning in large language models. *Preprint*, arXiv:2602.02462.

Farzaneh Bayan Memar, Hanneke Huls, and Matthijs ten Hove. 2026. Ellat at semeval-2026 task 11: Comparing encoder and decoder models for syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

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.

Wicaksono Leksono Muhamad, Joanito Agili Lopo, Tack Hwa Wong, Muhammad Ravi Shulthan Habibi, and Samuel Cahyawijaya. 2026. Itlc at semeval-2026 task 11: Normalization and deterministic parsing for formal reasoning in llms. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

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.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Wiebke Petersen, Cherine Jaziri, and Diem-Xuan Tran. 2026. HHU-SyLo at SemEval-2026 task 11: Logic in the loop – hybridizing LLMs and theorem provers for robust formal reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Xin Quan, Marco Valentino, Louise Dennis, and André 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*, pages 2933–2958.

Rakshith R and Ankush Chopra. 2026. Aicoe-tredence at semeval-2026 task 11: Mitigating content bias in syllogisms via symbolic logic-language decoupling. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

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)*, pages 17222–17240, Vienna, Austria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

S Seals and Valerie Shalin. 2024. Evaluating the deductive competence of large 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 8606–8622.

Soumyajit Roy Setu and Manav Malhotra. 2026. Modusponens at semeval-2026 task 11: Breaking the plausibility trap in llms via conflict-aware ensembling. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Abdullah Shaikh, Zain Naqi, Taha Zahid, Sandesh Kumar, and Abdul Samad. 2026. HABIB_TAZ at SemEval-2026 task 11: Disentangling formal logic from content via synthetic training and multi-objective optimization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Lakksh Sharma, Jatin Bedi, and Krish Sharma. 2026a. Lakksh at semeval-2026 task 11(1 & 2): Neuro-symbolic decomposition to mitigate content bias in syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Lakksh Sharma, Rhea Singhal, Krish Sharma, and Jatin Bedi. 2026b. Lakksh at semeval-2026 task 11(3&4): Instability-triggered compute for multilingual syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Peiyang Song, Pengrui Han, and Noah Goodman. 2026. Large language model reasoning failures. *Transactions on Machine Learning Research*. Survey Certification.

C. Sai Aravind Sunil Saumya and C. Pothan Sai Reddy. 2026. SpyComet at SemEval-2026 task 11: When adversarial debiasing backfires - a comparison of data-level and model-level debiasing. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Hidetsune Takahashi. 2026. Hidetsune at SemEval-2026 task 11: Adapting pretrained reasoning models with deep supervision and inference refinement for content-independent validity classification. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Xingwei Tan, Marco Valentino, Mahmud Elahi Akhter, Maria Liakata, and Nikolaos Aletras. 2025. Enhancing logical reasoning in language models via symbolically-guided Monte Carlo process supervision. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31886–31900, Suzhou, China. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

P. K. Thiyagarajaa and Durairaj Thenmozhi. 2026. Thiyaga6851 at semeval-2026 task 11: Disentangling content and formal reasoning in large language models using neuro-symbolic mapping. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Chi-Nguyen Tran, Duy Minh Dao Sy, Trung Kiet Huynh, Phu-Hoa Pham, and Lam Phu Quy Nguyen. 2026. HCMUS_DroneBoys at SemEval-2026 task 11: Asymmetric counterfactual debiasing and rank-sensitive logical invariance adaptation for syllogistic reasoning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Noah Tratzsch, Asmaa Al-Raian, Mounika Marreddy, and Alexander Mehler. 2026. Semeval-2026 task 11: Reducing content effects using layered activation steering. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2026. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

Yu Wang, You Zhang, Hao Zhang, and Dan Xu. 2026. Ynu-nlp at semeval-2026 task 11: A neuro-symbolic approach with reflexion mechanism disentangling content and formal reasoning in language models. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2024. Syllobio-nli: Evaluating large language models on biomedical syllogistic reasoning. *arXiv preprint arXiv:2410.14399*.

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

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao,

and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Syllogistic Schemes

Here is the list of the 24 valid syllogistic schemes adopted to construct the dataset:

```
Figure 1

Schema: Barbara (AAA-1)
Premise 1: All <B> are <A>
Premise 2: All <C> are <B>
Conclusions: All <C> are <A>

Schema: Barbari (AAI-1)
Premise 1: All <B> are <A>
Premise 2: All <C> are <B>
Conclusions: Some <C> are <A>

Schema: Celarent (EAE-1)
Premise 1: No <B> are <A>
Premise 2: All <C> are <B>
Conclusions: No <C> are <A>

Schema: Celaront (EAO-1)
Premise 1: No <B> are <A>
Premise 2: All <C> are <B>
Conclusions: Some <C> are not <A>

Schema: Darii (AII-1)
Premise 1: All <B> are <A>
Premise 2: Some <C> are <B>
Conclusions: Some <C> are <A>

Schema: Ferio (EIO-1)
Premise 1: No <B> are <A>
Premise 2: Some <C> are <B>
Conclusions: Some <C> are not <A>

Figure 2

Schema: Camestres (AEE-2)
Premise 1: All <A> are <B>
Premise 2: No <C> are <B>
Conclusions: No <C> are <A>

Schema: Camestros (AEO-2)
Premise 1: All <A> are <B>
Premise 2: No <C> are <B>
Conclusions: Some <C> are not <A>

Schema: Cesare (EAE-2)
Premise 1: No <A> are <B>
Premise 2: All <C> are <B>
Conclusions: No <C> are <A>

Schema: Cesaro (EAO-2)
Premise 1: No <A> are <B>
Premise 2: All <C> are <B>
Conclusions: Some <C> are not <A>

Schema: Baroco (AOO-2)
Premise 1: All <A> are <B>
Premise 2: Some <C> are not <B>
Conclusions: Some <C> are not <A>

Schema: Festino (EIO-2)
Premise 1: No <A> are <B>
Premise 2: Some <C> are <B>
Conclusions: Some <C> are not <A>

Figure 3

Schema: Darapti (AAI-3)
Premise 1: All <B> are <A>
Premise 2: All <B> are <C>
Conclusions: Some <C> are <A>

Schema: Datisi (AII-3)
Premise 1: All <B> are <A>
Premise 2: Some <B> are <C>
Conclusions: Some <C> are <A>

Schema: Disamis (IAI-3)
Premise 1: Some <B> are <A>
Premise 2: All <B> are <C>
Conclusions: Some <C> are <A>

Schema: Felapton (EAO-3)
Premise 1: No <B> are <A>
Premise 2: All <B> are <C>
Conclusions: Some <C> are not <A>

Schema: Bocardo (OAO-3)
Premise 1: Some <B> are not <A>
Premise 2: All <B> are <C>
Conclusions: Some <C> are not <A>

Schema: Ferison (EIO-3)
Premise 1: No <B> are <A>
Premise 2: Some <B> are <C>
Conclusions: Some <C> are not <A>

Figure 4

Schema: Bamalip (AAI-4)
Premise 1: All <A> are <B>
Premise 2: All <B> are <C>
Conclusions: Some <C> are <A>

Schema: Camenes (AEE-4)
Premise 1: All <A> are <B>
Premise 2: No <B> are <C>
Conclusions: No <C> are <A>

Schema: Camenos (AEO-4)
Premise 1: All <A> are <B>
Premise 2: No <B> are <C>
Conclusions: Some <C> are not <A>

Schema: Dimaris (IAI-4)
Premise 1: Some <A> are <B>
Premise 2: All <B> are <C>
Conclusions: Some <C> are <A>

Schema: Fesapo (EAO-4)
Premise 1: No <A> are <B>
Premise 2: All <B> are <C>
Conclusions: Some <C> are not <A>

Schema: Fresison (EIO-4)
```

```
Premise 1: No <A> are <B>
Premise 2: Some <B> are <C>
Conclusions: Some <C> are not <A>
```

For invalid syllogisms, we construct examples of logical fallacies instantiating the following templates:

```
Schema: Undistributed Middle (A-A-A
    variation)
Premise 1: All <A> are <B>
Premise 2: All <C> are <B>
Conclusions: All <C> are <A> | No <C>
    are <A> | Some <C> are <A>

Schema: Undistributed Middle (A-I-I
    variation)
Premise 1: All <A> are <B>
Premise 2: Some <C> are <B>
Conclusions: Some <C> are <A> | All <C>
    are <A>

Schema: Illicit Major (A-E-E variation)
Premise 1: All <B> are <A>
Premise 2: No <C> are <B>
Conclusions: No <C> are <A>

Schema: Illicit Major (A-O-O variation)
Premise 1: All <B> are <A>
Premise 2: Some <C> are not <B>
Conclusions: Some <C> are not <A>

Schema: Illicit Minor (A-A-A variation)
Premise 1: All <B> are <A>
Premise 2: All <B> are <C>
Conclusions: All <C> are <A>

Schema: Illicit Minor (I-A-A variation)
Premise 1: Some <A> are <B>
Premise 2: All <B> are <C>
Conclusions: All <C> are <A>

Schema: Affirmative Conclusion from
    Negative Premise (E-A-A variation)
Premise 1: No <A> are <B>
Premise 2: All <C> are <B>
Conclusions: All <C> are <A>

Schema: Affirmative Conclusion from
    Negative Premise (E-A-I variation)
Premise 1: No <B> are <A>
Premise 2: All <C> are <B>
Conclusions: Some <C> are <A>

Schema: Affirmative Conclusion from
    Negative Premise (O-A-I variation)
Premise 1: Some <B> are not <A>
Premise 2: All <B> are <C>
Conclusions: Some <C> are <A>

Schema: Two Negative Premises (E-E-E
    variation)
Premise 1: No <A> are <B>
Premise 2: No <C> are <B>
Conclusions: No <C> are <A> | Some <C>
    are <A>

Schema: Two Negative Premises (E-E-O
    variation)
Premise 1: No <B> are <A>
Premise 2: No <B> are <C>
Conclusions: Some <C> are not <A> | No <
    C> are <A>

Schema: Two Negative Premises (E-O-O
    variation)
Premise 1: No <A> are <B>
Premise 2: Some <C> are not <B>
Conclusions: Some <C> are not <A>
```