# 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

**Rongchuan Luo, Jin Wang, and Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: {luorongchuan, wangjin, xjzhang}@ynu.edu.cn

## Abstract

This paper addresses SemEval-2026 Task 11, which focused on mitigating content effects in syllogistic reasoning. Logical validity is often conflated with semantic plausibility in large language models. Prior methods rely on standard fine-tuning or prompting, without explicit bias control. A rule- and template-based symbolic data augmentation framework is proposed for fine-tuning the `Qwen2-1.5B-Instruct` model and instruction-tuning the `XLM-RoBERTa-large` model. Logic-preserving synthetic data are generated through lexical rules. The system is ranked 1st in Task 1 with a perfect overall score of 100, and 6th in Task 3 with a score of 56.97. Code is publicly available at: https://github.com/YNU-HPCC/semeval-2026-task11.

## 1 Introduction

Syllogistic reasoning, a fundamental deductive process deriving logically valid conclusions from premises based on formal rules (Nie et al., 2020), serves as a standard evaluation paradigm for natural language inference (NLI) models due to its rigid, rule-based structure (Wu et al., 2023; Rudinger et al., 2020).

As large language models (LLMs) advance across diverse NLP tasks (Grattafiori et al., 2024; Achiam et al., 2023; Wei et al., 2022), evaluating their emergent reasoning abilities has become a critical research focus (Clark et al., 2020). While some studies suggest LLMs implicitly develop formal reasoning skills during pretraining (Dai et al., 2022; Kojima et al., 2022), their actual logical competence remains inconsistent. Despite robust commonsense reasoning capabilities, LLM performance on formal logical tasks is frequently undermined by a cognitive bias known as *content effects*, where semantic believability interferes with strict logical evaluation (Bertolazzi et al., 2024; Khemlani and Johnson-Laird, 2012).

To address this vulnerability, we propose a bias-suppression fine-tuning framework for Subtasks 1 and 3 of SemEval-2026 Task 11, utilizing `Qwen2-1.5B-Instruct` (Bai et al., 2023) and `XLM-RoBERTa-large`. Our approach explicitly decouples semantic plausibility from logical validity during training. To support this, we build upon prior methodologies (Bertolazzi et al., 2024; Kim et al., 2025; Wysocka et al., 2025; Valentino et al., 2025) to construct a synthetic dataset of over 20,000 categorical syllogisms across 64 abstract patterns. This augmented data systematically balances the intersections of valid/invalid and plausible/implausible arguments.

Empirical results demonstrate the effectiveness of this decoupling strategy. Our system ranked 1st in the English syllogistic validity classification task (Subtask 1) with a perfect comprehensive score of 100, and 6th in the multilingual setting (Subtask 3) with a score of 56.97.

## 2 Related work

Logical inference is frequently disrupted by semantic believability, a cognitive shortcut known as belief bias or the content effect (Evans et al., 1983). Large language models (LLMs) inherit this flaw, prioritizing pretraining semantic priors over abstract logical rules in deductive tasks like syllogistic reasoning (Dasgupta et al., 2022). Consequently, models frequently endorse invalid but factually plausible conclusions (Ozeki et al., 2024; Eisape et al., 2024). To quantify this, specialized datasets such as NEUBAROCO (Ozeki et al., 2024) and SYLLOBIO-NLI (Wysocka et al., 2025) isolate content-driven errors by explicitly contrasting belief-consistent, belief-violating, and structurally arbitrary syllogisms (Eisape et al., 2024).

Early mitigation efforts primarily relied on prompting strategies. While Chain-of-Thought (CoT) improves accuracy on valid arguments, it remains vulnerable to content effects, routinely fail-



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 277–283

ing on invalid yet plausible syllogisms (Lyu et al., 2023; Wang et al., 2022; Bertolazzi et al., 2024; Saparov and He, 2022). Seeking stricter logical adherence, researchers have integrated neural models with symbolic solvers (Quan et al., 2024) or introduced quasi-symbolic abstractions via templated argument parsing (Ranaldi et al., 2025). Though effective, these hybrid architectures heavily depend on external theorem provers or restricted structural formats, limiting their open-domain applicability.

At the model-parameter level, supervised fine-tuning (SFT) can reduce bias, but standard cross-entropy loss fails to distinguish between critical error boundaries, such as plausible-invalid versus implausible-valid cases (Ranaldi et al., 2025). Other advanced interventions include fine-grained activation steering to suppress bias-inducing layers without retraining (Valentino et al., 2025), and two-stage reinforcement learning frameworks like SYLER to enforce structural reasoning paths (Zhang et al., 2025). Despite their theoretical appeal, these techniques often require complex architectural interventions or pipeline engineering.

The persistence of content effects across both prompted and heavily intervened models underscores a fundamental limitation: architectural tweaks or prompting alone are insufficient. Instead, bias-aware decoupling mechanisms must be explicitly embedded into the training objective to reliably separate semantic plausibility from logical validity.

## 3 Methodology

### 3.1 Task Overview

This study is based on SemEval-2026 Task 11 (Valentino et al., 2026), focusing on two core subtasks: Subtask 1 (English syllogistic reasoning) and Subtask 3 (multilingual syllogistic reasoning). In Subtask 1, the validity of syllogisms presented in English is to be judged solely on formal logical grounds, without regard to whether the content aligns with real-world knowledge. Each input consists of a syllogism comprising two premises and a conclusion, and a binary label (valid or invalid) is to be assigned. Subtask 3 extends this setting to a multilingual context, covering twelve languages: English, German, Spanish, French, Italian, Dutch, Portuguese, Russian, Chinese, Swahili, Bengali, and Telugu. The objective remains the same as that of Subtask 1. Both subtasks are evaluated using a composite metric that combines accuracy and content effect, designed to assess the generalization and robustness of current large language models in formal logical reasoning across monolingual and multilingual settings.

### 3.2 A Rule-Based and Template-Driven Symbolic Data Augmentation Method

In the task, a syllogism $S$ is defined as a triple $S = (P_1, P_2, C)$, where $P_1$ and $P_2$ are two categorical premises and $C$ is the conclusion. They share the canonical structure *Quantifier X is/are Y*. As shown in Table 1 (left), four categorical *moods* can be exhibited by such statements. These *moods* are conventionally abbreviated by letters: affirmative forms as A and I, and negative forms as E and O (Bertolazzi et al., 2024).

The *figure* of a syllogism is defined by the order in which terms appear in the premises. Four *figures* are numbered from 1 to 4. Exactly one common term is shared by the two premises, denoted as $b$ in Table 2, while the other two distinct terms, $a$ and $c$ , are linked through logical inference to form the conclusion (Guzmán et al., 2024). Consequently, a predicate chain is constituted by the syllogism, through which the subject and predicate of the conclusion are connected via the shared middle term.

A syllogistic inference is fully determined by the combination of the *moods* of the two premises and the *figure*. For example, AE2 denotes a syllogism in which the first premise is of *mood* A, the second premise is of *mood* E, and the *figure* is 2.

Since each premise can instantiate any one of the four categorical *moods* (A, E, I, O), there are $4 \times 4 = 16$ possible *mood* pairs. Coupled with the four *figures*, this yields a total of $16 \times 4 = 64$ distinct *mood–figure* combinations (Bertolazzi et al., 2024). We classify 37 of these 64 combinations as invalid, meaning they do not logically entail any conclusion under classical syllogistic rules. Meanwhile, the remaining 27 are deemed valid, as each supports at least one logically derivable conclusion.

Following the above structure, a dataset comprising approximately 20,000 English syllogisms was constructed through data augmentation. All 64 canonical syllogistic forms are covered, and explicit control is exerted along two orthogonal dimensions: formal validity and semantic plausibility. Representative examples are provided in Table 2.

### 3.3 Qwen2-1.5B-Instruct for Syllogistic Reasoning in English

For Subtask 1, we build our system upon `Qwen2-1.5B-Instruct`. We first preprocess the

| *Moods* | | *Figures* | | | | *Schema: AE2* | |
|---|---|---|---|---|---|---|---|
| Affirmative | Negative | 1 | 2 | 3 | 4 | Pre/Con | Statement |
| A: All $a$ are $b$ | E: No $a$ are $b$ | P1: a–b | b–a | a–b | b–a | P1: | All $b$ are $a$ (A) |
| I: Some $a$ are $b$ | O: Some $a$ are not $b$ | P2: b–c | c–b | c–b | b–c | P2: | No $c$ are $b$ (E) |
| | | | | | | C: | Some $a$ are not $c$ |

Table 1: Basic Components of Syllogisms: *Moods*, *Figures*, and the *AE2 Schema*

| | |
|---|---|
| **Valid Plausible** | All siameses are cats. All cats are felines. All siameses are felines. |
| **Valid Implausible** | All felines are cats. All cats are dogs. All felines are dogs. |
| **Invalid Plausible** | All siameses are felines. All cats are felines. All siameses are cats |
| **Invalid Implausible** | All siameses are felines. All cats are felines. All cats are siameses. |

Table 2: Examples of the Four Types of Syllogisms

raw syllogisms by segmenting them into distinct premises and conclusions, explicitly retaining both logical validity and semantic plausibility labels for downstream bias evaluation.

Because decoder-only models like Qwen2 lack a dedicated classification token (e.g., [CLS]), we extract the hidden state of the last non-padding token to serve as the sequence-level representation. This vector feeds into a lightweight classification head comprising a dropout layer and a linear projection, allowing us to optimize the entire pipeline end-to-end using cross-entropy loss.

During supervised fine-tuning, we apply stratified sampling to guarantee consistent joint distributions of validity and plausibility across the training and validation splits. For evaluation, we supplement standard classification accuracy with specialized content-effect metrics, namely IPCE, CPCE, and TCE. These metrics specifically quantify the impact of semantic plausibility on model predictions, providing a more rigorous and comprehensive assessment of genuine logical reasoning capabilities.Figure 1 illustrates our end-to-end data augmentation and training pipeline.

### 3.4 XLM-RoBERTa-large for Multilingual Syllogistic Reasoning

`XLM-RoBERTa-large` is adopted as the base model for Subtask 3. This model is a multilingual pretrained language model based on the Transformer encoder architecture, supporting over 100 languages and demonstrating strong cross-lingual semantic understanding. The system pipeline is structured as follows. First, the original syllogistic text is robustly segmented to extract Premise 1, Premise 2, and the Conclusion. These components are then combined into a structured natural language instruction:

> Determine if the conclusion LOGICALLY FOLLOWS from the premises, regardless of whether it sounds plausible or true in the real world. Answer only based on formal reasoning.

Under this instructional format, `XLM-RoBERTa-large` is fine-tuned via instruction-based supervised fine-tuning. The `[CLS]` token is used as input to the classification head for binary classification of logical validity. During training, stratified sampling is applied based on the joint distribution of validity and plausibility labels to partition the dataset into training and validation subsets, ensuring balanced representation across all combinations of semantic plausibility and logical validity, thereby mitigating potential biases.

## 4 Experimental Details

**Dataset.** Our experiments build upon the official SemEval-2026 Task 11 dataset, supplemented by synthetic data augmentation. The original release provides 1,040 manually annotated syllogisms, each featuring two premises and a conclusion explicitly labeled for logical validity and semantic plausibility.

Given that this initial data scale is insufficient for robust LLM fine-tuning, we generated approximately 20,000 additional high-quality instances.

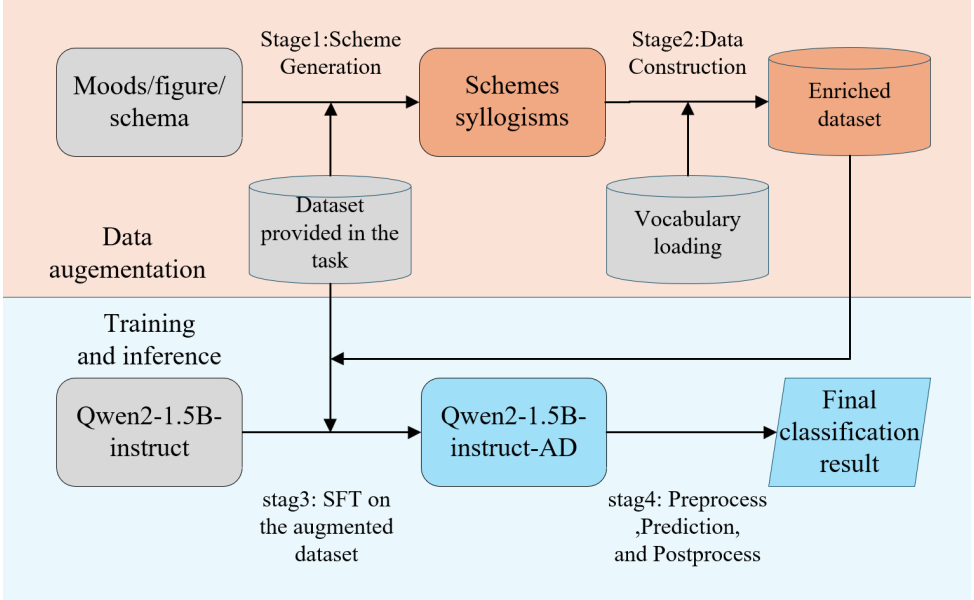


Figure 1: Data Augmentation and Training

| Augmentation | Method | Metrics | | |
|---|---|---|---|---|
| | | Combined Score | Validity Acc. | Content Effect |
| no aug. | xlm-roberta-large | 18.29 | 73.30 | 19.24 |
| | bart-large-mnli | 31.60 | 94.76 | 6.38 |
| | deberta-v3-large | 20.36 | 71.73 | 11.46 |
| | Qwen2.5-1.5B-Instruct | 33.69 | 95.81 | 5.32 |
| | Qwen2-1.5B-Instruct | 25.52 | 85.86 | 9.64 |
| aug. | xlm-roberta-large | 26.00 | 78.53 | 6.54 |
| | bart-large-mnli | 40.45 | 98.43 | 3.19 |
| | deberta-v3-large | 40.54 | 98.43 | 3.17 |
| | Qwen2.5-1.5B-Instruct | 45.99 | 98.43 | 2.13 |
| | Qwen2-1.5B-Instruct | **100.00** | **100.00** | **0.00** |

Table 3: Performance comparison on Task 1 with and without data augmentation.

| Augmentation | Method | Metrics | | |
|---|---|---|---|---|
| | | Combined Score | Validity Acc. | Content Effect |
| no aug. | berta-base | 23.55 | 78.65 | 9.38 |
| | mDeBERTa-v3-base-mnli | 29.24 | 88.54 | 6.60 |
| | Qwen2.5-1.5B-Instruct | 24.87 | 87.50 | 11.41 |
| | Qwen2-1.5B-Instruct | 24.72 | 89.06 | 12.50 |
| | xlm-roberta-large | 25.37 | 71.88 | 5.26 |
| aug. | berta-base | 28.74 | 90.62 | 7.61 |
| | mDeBERTa-v3-base-mnli | 33.54 | 94.79 | 5.21 |
| | Qwen2.5-1.5B-Instruct | 37.51 | 87.50 | 2.79 |
| | Qwen2-1.5B-Instruct | 49.82 | 90.62 | 1.27 |
| | xlm-roberta-large | **56.97** | **96.35** | **1.00** |

Table 4: Performance comparison on Task 3 with and without data augmentation.

Leveraging WordNet (Miller, 1995) as an external knowledge base, we populated the established logical templates while strictly enforcing the original semantic constraints. Our generation pipeline meticulously preserves the formal structure and semantic distribution of the source data. To guarantee data integrity, we manually verified the labels across all synthetic samples, ultimately yielding a combined training corpus of roughly 21,040 instances.

**Evaluation Metric.** To evaluate both reasoning accuracy and robustness against content bias, the task employs a composite score:

$$\text{Score} = \frac{\text{Acc}}{1 + \ln(1 + \text{CE})} \tag{1}$$

where Acc is the logical validity classification accuracy, and CE (Total Content Effect (Valentino et al., 2025)) measures the influence of semantic plausibility. CE averages the Intra-Plausibility ($B_{\text{intra}}$) and Cross-Plausibility ($B_{\text{inter}}$) biases:

$$B_{\text{intra}} = \frac{1}{2}\left(|A_{11} - A_{10}| + |A_{01} - A_{00}|\right) \tag{2}$$

$$B_{\text{inter}} = \frac{1}{2}\left(|A_{11} - A_{01}| + |A_{10} - A_{00}|\right) \tag{3}$$

$$\text{CE} = \frac{B_{\text{intra}} + B_{\text{inter}}}{2} \tag{4}$$

Here, $A_{ab}$ $(a, b \in \{0, 1\})$ represents the accuracy on the subset with validity $a$ and plausibility $b$. $B_{\text{intra}}$ and $B_{\text{inter}}$ capture the model's performance variance under fixed validity and fixed plausibility conditions, respectively.

**Implementation Details.** In the final submission, `Qwen/Qwen2-1.5B-Instruct` is used for Task 1 and `FacebookAI/xlm-roberta-large` is used for Task 3 to evaluate performance on logical validity judgment in English and multilingual syllogisms in SemEval-2026. Input is formatted as a single sequence in both tasks. A concise template is employed in Task 1. In Task 3, a strong logical instruction is added to suppress bias from semantic plausibility. Tokenization is performed using each models official tokenizer. The maximum sequence length is set to 512. For `Qwen2-1.5B-Instruct`, a linear classification head with dropout is added on top of the last hidden state, as it is a decoder-only architecture. For `xlm-roberta-large`, the built-in sequence classification head is fine-tuned. Fine-tuning is conducted on augmented datasets. Stratified splits are applied to maintain balanced distributions across the four validity–plausibility categories.

**Parameters Fine-tuning.** In the Dev set, the `AdamW` optimizer is used with a learning rate of $1 \times 10^{-5}$ and a batch size of 16.

**Comparative Results.** Tables 3 and 4 demonstrate the empirical superiority of our integrated methodology. For English syllogisms (Task 1), our system secured the 1st place ranking with a perfect combined score of 100.00. This reflects a 100% validity accuracy and a complete elimination of the content effect (CE = 0.00), confirming the model's strict adherence to formal logic. In the multilingual track (Task 3), the framework achieved a combined score of 56.97 (ranking 6th), driven by a high validity accuracy of 96.35% and a heavily suppressed content effect of 1.00.

## 5 Conclusions

This paper presented a comprehensive fine-tuning framework to tackle content effects in LLM syllogistic reasoning. By synergizing data augmentation, stratified resampling, and instruction tuning, we successfully decoupled logical validity from semantic plausibility. The effectiveness of this architecture is highlighted by our top-tier performance in the SemEval-2026 evaluations, particularly achieving absolute robustness against belief bias in the monolingual setting. Moving forward, we plan to refine the joint prediction mechanics for validity and plausibility, with a particular focus on resolving complex cross-lingual semantic conflicts in multilingual environments.

## Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051, the Scientific Research and Innovation Project of Postgraduate Students in the Academic Degree of YunNan University (Nos. KC-252513686), and the Project of Yunnan Provincial Department of Education Science Research (Nos. 2026Y0187). The authors would like to thank the anonymous reviewers for their constructive comments.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

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

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

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*.

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

J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

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*.

Manuel Vargas Guzmán, Jakub Szymanik, and Maciej Malicki. 2024. Testing the limits of logical reasoning in neural and hybrid models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2267–2279.

Sangeet Khemlani and Philip N Johnson-Laird. 2012. Theories of the syllogism: A meta-analysis. *Psychological bulletin*, 138(3):427.

Geonhee Kim, Marco Valentino, and André 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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)*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

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. *arXiv preprint arXiv:2408.04403*.

Xin Quan, Marco Valentino, Louise A Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through llm-symbolic theorem proving. *arXiv preprint arXiv:2405.01379*.

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.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

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

Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and Andre 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)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Yongkang Wu, Meng Han, Yutao Zhu, Lei Li, Xinyu Zhang, Ruofei Lai, Xiaoguang Li, Yuanhang Ren, Zhicheng Dou, and Zhao Cao. 2023. Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2347–2367.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2025. Syllobionli: 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.

Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025. Syler: A framework for explicit syllogistic legal reasoning in large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4117–4127.