SEF-CLGC at SemEval-2026 Task 11: Logical Notation Impact on Language Model Performance

Hanna Abi Akl[1,2], Fabien Gandon[1], Catherine Faron[1], Pierre Monnin[1]
[1]Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France
[2]Data ScienceTech Institute, Paris, France
hanna.abi-akl@inria.fr

Abstract

This paper revisits our pipeline called Syllogistic Evaluation Framework-Common Logic Grammar Construction (SEF-CLGC). We combine formal logical notations with Small Language Models (SLMs) to evaluate reasoning performance on the SemEval-2026 Task 11 Subtask 1: Disentangling Content and Formal Reasoning in Large Language Models. Our experiments show that by relying solely on SLMs, trained on a combination of natural and symbolic languages, our best model achieves a content score of 27.80% on the task while significantly lowering the content bias in reasoning.

1 Introduction

Against the trend of scaling bigger language models (LMs), Small Language Models (SLMs) have re-emerged as powerful agents capable of performing complex tasks like reasoning in many domains (Masri et al., 2026; Srivastava et al., 2025). This emergence has led to benchmarking SLMs across a variety of reasoning tasks like mathematics and common sense problems (Zhuang et al., 2025). Further exploration has led to investigating neural-based (Wang et al., 2025; Kim et al., 2025b) and neuro-symbolic (Lyu et al., 2023; Quan et al., 2024; Han et al., 2025) techniques to enhance their reasoning abilities. The SemEval-2026 Task 11 Subtask 1 (Valentino et al., 2026) challenges LM reasoning on determining the validity of syllogisms. We leverage the Syllogistic Evaluation Framework-Common Logic Grammar Construction (SEF-CLGC) pipeline previously introduced in (Akl, 2025) to test the reasoning capabilities of SLMs by training them on different logical notations inspired from formal Knowledge Representation (KR) languages and assessing their performance on the challenge.

ID	Syllogism	Validity	Plausibility
50146f21-d265-4e3a-8d93-8165cdbe89a3	All cars are a type of vehicle. No animal is a car. Therefore, no animal can be a vehicle.	False	True
08408587-3887-4246-9d6f-7a4492ad48c7	Anyone who is a rose is red. Some flowers are not red. From this, all flowers are roses.	False	False

Table 1: Example data.

2 Background

2.1 Task Description

The goal of Task 11 Subtask 1 is to determine the correct validity label (i.e. "true" or "false") of a given syllogism. A data point consists of a syllogism composed of premises and a conclusion in natural language (English) and associated information. Table 1 is an example. Plausibility indicates whether the arguments of a syllogism are aligned (i.e. "true") or misaligned (i.e. "false") with real-world knowledge. This subtask has 2 phases: a Training phase to prepare and train participating systems on pilot (80 syllogisms) and training (960 syllogisms) sets, and the official Evaluation phase where systems are evaluated and ranked on a blind evaluation set of 191 syllogisms.

2.2 Related Work

Different systems have been proposed in the literature to predict the validity of a syllogism. Works leveraging small and large LMs (Eisape et al., 2024; Ozeki et al., 2024) show that larger models make less mistakes but are still prone to the same reasoning biases (e.g. syllogistic fallacies) as humans. Other works (Dasgupta et al., 2022; Bertolazzi et al., 2024) compare Supervised Fine-Tuning (SFT) and In-Context Learning (ICL) strategies and show that SFT has a better mitigating effect than ICL on bias on small and mid-sized LMs. Reasoning limitations in LMs paved the way for neuro-symbolic systems that integrate rules (Seals and

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2251–2261

Shalin, 2024; Valentino et al., 2025; Wysocka et al., 2025) in the model prompt to control bias and enhance performance. Other areas of research leverage prompting techniques like Chain-Of-Thought (COT) and explanation generation as a method of auto-correction and auto-evaluation of LM reasoning (Xu et al., 2024). Lastly, multi-stage systems that translate syllogisms from Natural Language (NL) to formal languages like First-Order Logic (FOL) and combine natural explanations with theorem-proving (Ranaldi et al., 2025; Kim et al., 2025a; Maraia et al., 2026) show boosted performance in predicting syllogism validity. Our work follows in that direction.

3 System Architecture

3.1 Enhanced SEF-CLGC

We reuse the SEF-CLGC pipeline from (Akl, 2025) which transforms syllogisms in FOL notation into other logical notations: CLIF, CGIF and TFL+. We enhance the pipeline by introducing two new notations: CLINGO and a custom MINIFOL2.

CLINGO (Gebser et al., 2014) is a language for Answer Set Programming (ASP), a form of declarative logic programming used to model and solve combinatorial search problems.

MINIFOL2 spans from the custom MINIFOL notation introduced in (Akl, 2025) which replaces FOL operators with their Boolean equivalents (e.g. "∧" with "&"). MINIFOL2 also eliminates the "∃" quantifier from the syntax. The notation is used as a baseline to study the effects of slight syntactic changes on LMs.

3.2 Pipeline

The methodology pipeline is shown in Figure 1.

NL-FOL Translation: The starting point for SEF-CLGC is FOL notation. Since the Subtask data is provided in NL, we first translate the syllogisms into FOL notation using OpenAI's ChatGPT 5.2 model[1]. We chose this model because preliminary tests showed good performance but other models will be tested in extensions of this work. Translations are validated manually on a random 20% sample from the training set due to its size as well as the entire evaluation set from the evaluation phase.

SEF-CLGC: The dataset in FOL notation is then given to the SEF-CLGC framework. The SEF component categorizes syllogisms into 4 categories as per (Akl, 2025): Hypothetical (i.e. containing an implication), Disjunctive (i.e. containing a disjunction), Categorical (i.e. any syllogism of 2 premises and a conclusion not belonging to the aforementioned categories) and Complex (i.e. not belonging to any other category). Appendix A.1 shows examples of SEF categories. The CLGC component uses the Backus-Naur Form (BNF) grammar of logical notations including FOL to generate the Abstract Syntax Tree (AST) of each syllogism and validate it using a syntactic parser. Each syllogism is then transformed to the target logical notation from the AST and the BNF grammar of that notation. The resulting dataset contains all initial syllogisms along with their transformations in each of the logical notations (i.e. KR notations in Figure 1). Appendix A.1 also shows examples of the transformation from FOL to logical notations.

Validity Prediction: Syllogisms in a notation (e.g. NL, FOL) or a combination of notations (e.g. NL-FOL, FOL-CLIF-CGIF) are given with their validity labels as input in SFT to LMs for training and evaluation during the Training phase. For the blind Evaluation, only the syllogisms are given to the LMs to predict the validity labels. Based on previous work (Han et al., 2024; Akl, 2025), we limit ourselves to very small LMs (i.e. less than 1 billion parameters) for their frugality and good results on this type of task.

4 Experimental Setup

We combine the official task pilot and training datasets as our working set and split it into train/validation/test sets to fine-tune our models. These splits are frozen and used for all our experiments. We use Google's Flan-T5-small[2] and large[3] and fine-tune them on Google Cloud T4 and A100 GPUs respectively. Inference is performed on an A100 GPU.

4.1 SEMEVAL Models

SEMEVAL Models are all vanilla Flan-T5 models directly trained on the working dataset splits. All models are trained on 5 epochs with a learning rate of 10^{-5} and a batch size of 4. All other parameters are set to default.

[1] https://chatgpt.com/share/699c606c-38c0-800e-ac41-6a55c246dd57

[2] https://huggingface.co/google/flan-t5-small

[3] https://huggingface.co/google/flan-t5-large

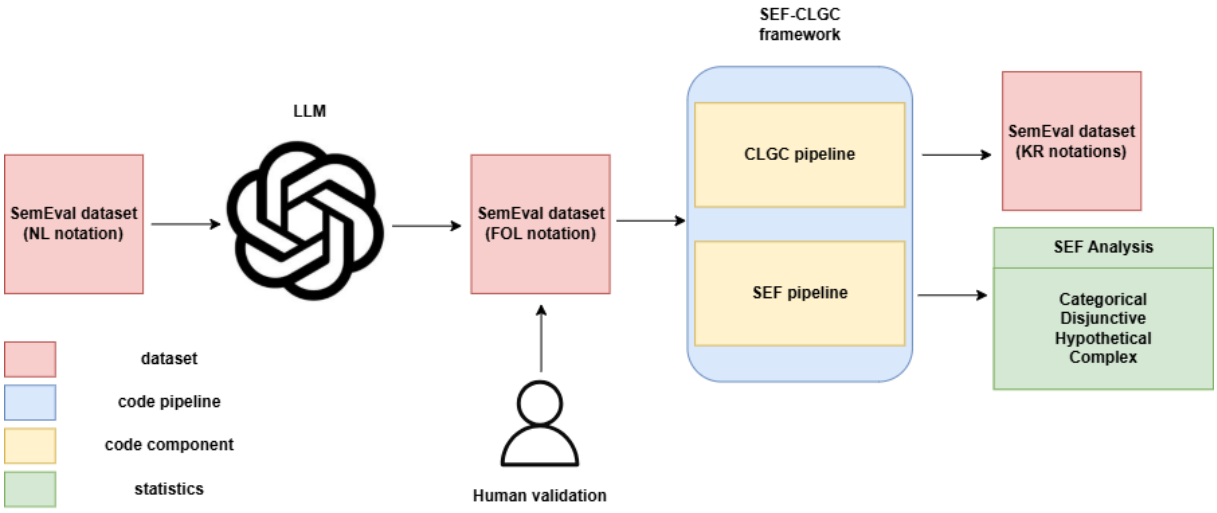

Figure 1: Dataset generation workflow.

FOLIO-SEMEVAL Models are Flan-T5 models that have already been fine-tuned on the FOLIO dataset as per (Akl, 2025) with the same parameters as SEMEVAL Models for epochs, learning rate and batch size and the rest being default. The models are then also fine-tuned on the working dataset, again under the same conditions as the SEMEVAL models.

4.2 Evaluation

All models receive the pair (syllogism, validity label) in one or more notation as input. Evaluation is based on the Content Score (CS):

$$CS = \frac{ACC}{1 + \log(1 + CE)} \quad (1)$$

where ACC is the overall accuracy and CE is the Content Effect. Since a syllogism can have one of two validity values (i.e. Valid or Invalid) for each of its two plausibility values (i.e. Plausible or Implausible), the CE is the average accuracy difference between Plausible (i.e. Plausible_Valid + Plausible_Invalid) and Implausible (i.e. Implausible_Valid + Implausible_Invalid) syllogisms.

5 Results

We present our training and evaluation results for both SEMEVAL and FOLIO-SEMEVAL models. Training results are scored on Accuracy only since the CE and CS calculations were performed on the official evaluation set.

5.1 SEMEVAL Models

Table 2 shows the training results for the SEMEVAL models. NL Flan-T5-small is used as a baseline performance for very small language models. Overall, of the Flan-T5-large models, NL performs best as it is the most seen notation for these models, followed closely by the NL-FOL and NL-CLIF notations. Adding more notations to the input does not seem to boost performance as shown by NL-FOL-CLIF which fails to beat NL-FOL or NL-CLIF. Table 3 shows the official results for this family. NL retains top spot, followed by NL-FOL since these two notations are the most widely seen in the pre-training of these models among the notations used for the task. NL-CLIF is bested by NL-FOL-CLIF possibly boosted by FOL and NL-CLINGO since CLINGO is closer to FOL notation. MINIFOL2 performs badly in both cases since it combines FOL syntax with Boolean operators that LMs are not accustomed to seeing together which may explain why it breaks down. TFL+ proves to be too abstract and cannot be learned efficiently.

5.2 FOLIO-SEMEVAL Models

Table 4 shows the training results for the FOLIO-SEMEVAL models. While NL-CLIF could not beat NL in the SEMEVAL models, its FOLIO-SEMEVAL counterpart having already been fine-tuned on FOLIO outperforms it and even beats NL. This suggests that the model can learn CLIF well, making it a reliable asset to solve syllogistic problems and underlining the added value of enriching

Notation	Acc	Pr	Re	F1
NL	**0.92**	**0.92**	**0.92**	**0.92**
NL-FOL	0.91	0.91	0.91	0.91
NL-CLIF	0.88	0.88	0.88	0.88
NL-FOL-CLIF	0.88	0.89	0.88	0.88
NL-CLINGO	0.87	0.87	0.87	0.87
FOL	0.80	0.81	0.80	0.80
*NL	0.78	0.79	0.78	0.78
CLINGO	0.77	0.81	0.77	0.77
CGIF	0.75	0.76	0.75	0.75
CLIF	0.74	0.75	0.74	0.73
MINIFOL2	0.72	0.72	0.72	0.71
FOL-CLIF-CLINGO	0.70	0.72	0.70	0.69
TFL+	0.61	0.61	0.61	0.61

Table 2: SEMEVAL Flan-T5-large training results: best results in bold and second-best underlined. Acc = Accuracy; Pr = Precision; Re = Recall; F1 = Weighted F1. * Flan-T5-small is used here as a baseline.

Notation	Acc	CE	CS
NL	**90.05**	**9.57**	**26.81**
NL-FOL	89.00	10.68	25.73
NL-FOL-CLIF	84.29	10.68	24.37
NL-CLINGO	84.29	10.70	24.36
NL-CLIF	83.76	10.70	24.21
CLINGO	67.53	23.95	16.01
CGIF	66.49	28.12	15.21
CLIF	65.96	29.16	14.96
MINIFOL2	64.92	23.95	15.39
FOL-CLIF-CLINGO	64.39	27.08	14.85
FOL	61.25	31.25	13.69
TFL+	59.16	5.89	20.18

Table 3: SEMEVAL Flan-T5-large evaluation results: best results in bold and second-best underlined. Acc = Accuracy; CE = Content Effect; CS = Combined Score.

Notation	Acc	Pr	Re	F1
NL-CLIF	**0.95**	**0.95**	**0.95**	**0.95**
NL	0.93	0.93	0.93	0.93
NL-FOL	0.92	0.92	0.92	0.92
CLIF	0.85	0.85	0.85	0.85
CLINGO	0.84	0.84	0.84	0.84
FOL	0.81	0.81	0.81	0.81
CGIF	0.75	0.76	0.75	0.75
MINIFOL2	0.75	0.75	0.75	0.74
TFL+	0.55	0.56	0.55	0.54

Table 4: FOLIO-SEMEVAL Flan-T5-large training results: best results in bold and second-best underlined.

Notation	Acc	CE	CS
NL-FOL	90.57	8.55	**27.80**
NL	**93.19**	9.57	27.74
NL-CLIF	89.00	13.85	24.06
CLIF	80.00	50.00	16.22
CLINGO	74.34	16.66	19.20
CGIF	69.10	20.83	16.92
FOL	66.49	**3.50**	26.54
TFL+	58.11	10.41	16.91
*MINIFOL2	N/A	N/A	N/A

Table 5: FOLIO-SEMEVAL Flan-T5-large evaluation results: best results in bold and second-best underlined. * scoring timed out on the evaluation platform.

NL with a formal notation. This is further emphasized by the results of the CLIF and CLINGO models which perform very well having only been fine-tuned on these notations on the relatively small FOLIO dataset. NL-CLIF and NL-FOL show the importance of neuro-symbolic integration to augment logical performance for LMs. Table 5 shows a slight change in ranking while retaining the behavior observed in training: NL-FOL takes top spot as the combination of FOL with NL reduces CE and results in a better overall CS score than plain NL. NL-CLIF falls to third place but still performs decently. Possible explanations for the performance loss might be LM NL-FOL translation errors that propagate to other logical notations or the notable increase in CE observed for this particular notation which suggests it learns some forms of plausible syllogisms better than others. Notations like TFL+ and MINIFOL2 perform poorly due to the unfamiliarity of LMs with their syntax which can result in performance breakdown. Prior work shows that re-training LM tokenizers on these syntaxes slightly boosts performance at small scale but breaks down as models scale (Akl, 2025). It is also worth noting that more abstract notations (e.g. TFL+) significantly decrease CE at the expense of reduced accuracy, as do neuro-symbolic combinations of NL and formal notations (e.g. NL-FOL-CLIF). The results also clearly highlight the positive effect of SFT on prior logical datasets to boost logical reasoning in LMs.

5.3 Ranking

Our overall best FOLIO-SEMEVAL evaluation model (i.e. NL-FOL) ranks 10^{th} in Accuracy and 7^{th} in CE in Subtask 1. Considering we limited our experiments to SLMs, our results show that it

is possible to have a model that is both competitive and frugal. For reproducibility purposes, we openly released the weights of our best training model NL-CLIF[4] and evaluation model NL[5] (best Accuracy) and will soon release those for NL-FOL (best CS).

5.4 Analysis

5.4.1 Dataset

Appendix A.2 shows the SEF classes for the working and evaluation sets. Categorical syllogisms dominate which might explain the performance of the SEMEVAL models since these syllogisms share the same syntactic patterns. Conversely, it might also explain the CE drift in the FOLIO-SEMEVAL models that were pre-fine-tuned on a more diverse dataset of SEF classes. The results suggest further investigation is needed on a dataset with more balanced SEF classes to re-assess the performance of the two families.

5.4.2 Prediction Errors

In Appendix A.3.1, Figure 2 shows that the best 3 training and evaluation SEMEVAL models (i.e. NL, NL-CLIF and NL-FOL) have a high False Positive (FP) count and a lower False Negative (FN) count. In comparison, Figure 3 shows that for the same best FOLIO-SEMEVAL models the gap between FP and FN is wider as these models have very high FP and very low FN counts. This suggests that our models are very good at detecting valid syllogisms but prone to making mistakes when predicting invalid ones.

Future analysis could focus on evaluating the decision threshold of these models or the syntactic structure of invalid syllogisms to boost their performance. The greater difference between FP and FN for FOLIO-SEMEVAL models might also be due to their prior exposure to the FOLIO dataset composed of 460 valid versus 351 invalid syllogism which may create the observed bias.

For plausibility, all FOLIO-SEMEVAL models make more mistakes on plausible syllogisms and very few errors on implausible ones from Figure 3 in Appendix A.3.1. Implausible syllogisms are logically structured arguments whose premises are unlikely or contrary to common sense. This suggests that these models are capable of handling imagined reasoning scenarios that stray from the common world knowledge they have learned in pre-training. The same pattern is observed for SEMEVAL models in Figure 2 with a reduced difference between plausible and implausible errors. The errors made on plausible syllogisms may be due to their arguments being borrowed from other syllogisms resulting in unfamiliar constructions to the LMs and bad predictions. We cannot attribute with certainty the difference in plausibility errors between the 2 families of models to the FOLIO dataset as the information on the plausibility of syllogisms is not explicitly divulged but offers us another avenue to explore in future research.

Figures 4 and 5 in A.3.2 show the common FP and FN errors for the SEMEVAL and FOLIO-SEMEVAL models respectively. Comparing both figures shows that for FP the highest error count is the common error among all 3 models which suggests that they are likely to reason on invalid syllogisms in similar fashion. The same cannot be said for FN errors which suggests models are very sensitive to changes in input combinations when reasoning on valid syllogisms. The comparison also shows that for both families NL-CLIF has the highest counts of uncommon errors with the other models in most cases (11.1%, 41.2%, 29.4% and 33.3% respectively) suggesting that models with this combination of notations are more prone to different reasoning behaviors.

6 Conclusion

In this paper, we extended the SEF-CLFC pipeline to the SemEval-2026 Task 11 Subtask 1 dataset. Our results show that by combining NL and a logical notation, SLMs can achieve up to 90% accuracy in classifying syllogism validity on the evaluation set. Our method also demonstrates that expressing logical problems in formal notations can reduce content bias. The SemEval challenge opened new questions for us to explore in future work, most notably enriching under-represented SEF categories as well as studying the impact of including more implausible syllogisms on logical notations for SLM reasoning. Possible future work will also include performance comparisons against larger LMs.

7 Limitations

A limitation of our work in the context of this task is the reliance on a commercial model to translate from NL to FOL. Updates or changes on the model

[4] https://huggingface.co/HannaAbiAkl/LOGIC-NL-CLIF-Flan-T5-Large

[5] https://huggingface.co/HannaAbiAkl/LOGIC-NL-Flan-T5-Large

usage could affect the quality of translation and potentially alter subsequent results in the SEF-CLGC pipeline.

References

Hanna Abi Akl. 2025. Investigating Language Model Capabilities to Represent and Process Formal Knowledge: A Preliminary Study to Assist Ontology Engineering. In *Rules and Reasoning, 9th International Joint Conference, RuleML+RR 2025*, Instabul, Turkey.

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.

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.

Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. 2014. Clingo= asp+ control: Preliminary report. *arXiv preprint arXiv:1405.3694*.

Dongge Han, Menglin Xia, Daniel Madrigal Diaz, Samuel Kessler, Ankur Mallick, Xuchao Zhang, Mirian Del Carmen Hipolito Garcia, Jin Xu, Victor Rühle, and Saravan Rajmohan. 2025. Enhancing reasoning capabilities of small language models with blueprints and prompt template search. *arXiv preprint arXiv:2506.08669*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2024. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031.

Geonhee Kim, Marco Valentino, and Andre Freitas. 2025a. 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.

Yujin Kim, Euiin Yi, Minu Kim, Se-Young Yun, and Taehyeon Kim. 2025b. Guiding reasoning in small language models with llm assistance. *arXiv preprint arXiv:2504.09923*.

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 *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

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

Yahya Masri, Emily Ma, Zifu Wang, Joseph Rogers, and Chaowei Yang. 2026. Benchmarking small language models and small reasoning language models on system log severity classification. *arXiv preprint arXiv:2601.07790*.

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.

Xin Quan, Marco Valentino, Louise A 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.

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.

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 8614–8630.

Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. 2025. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*.

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*.

ID	Syllogism	Validity	Plausibility	SEF
951df8bb-e9dc-4272-9db7-92fe5d28d337	Anything that is a dog has fur. There are some poodles that are dogs. There are no poodles that do not have fur.	False	True	Categorical
4480e5d5-495a-4928-a420-a3c74b9268a9	Every single mammal is an animal. Each and every feline is an animal. This makes it true that every feline is a mammal.	False	True	Hypothetical

Table 6: Example data with SEF categories.

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)*.

Chenyu Wang, Zishen Wan, Hao Kang, Emma Chen, Zhiqiang Xie, Tushar Krishna, Vijay Janapa Reddi, and Yilun Du. 2025. Slm-mux: Orchestrating small language models for reasoning. *arXiv preprint arXiv:2510.05077*.

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.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

Xialie Zhuang, Peixian Ma, Zhikai Jia, Zheng Cao, and Shiwei Liu. 2025. A technical study into small reasoning language models. *arXiv e-prints*, pages arXiv–2506.

A Appendix

A.1 Dataset Examples

In this section, we provide a concrete example from the working dataset and the resulting transformations from the SEF-CLGC framework. Table 6 shows the SEF classification on examples from the working dataset. Table 7 shows the resulting CLGC transformation into different logical notations.

A.2 Dataset Statistics

Table 8 shows the SEF classification statistics on the working dataset splits and the evaluation set. Tables 9 and 10 show the validity and plausibility count distribution on the same sets.

A.3 Error Analysis

A.3.1 Quantitative Error Analysis

Figure 2 shows the FP and FN error distribution on the validity prediction for the 3 best SEMEVAL models as well as the plausibility of the respective syllogisms. Figure 3 shows the same analysis for the best 3 FOLIO-SEMEVAL models.

A.3.2 Qualitative Error Analysis

We present a sample of qualitative error analysis for our best SEMEVAL and FOLIO-SEMEVAL models. Figure 4 shows the count of common errors among the 3 best SEMEVAL models and Figure 5 mirrors the analysis for the best 3 FOLIO-SEMEVAL models.

Table 11 shows some common and uncommon errors made by the best 3 SEMEVAL models. Table 12 shows the same analysis for the best 3 FOLIO-SEMEVAL models.

ID	FOL	CLIF	CGIF	CLINGO	TFL+	MINIFOL2
951df8bb-e9dc-4272-9db7-92fe5d28d337	∀x (AnythingThat(x) → DogHasFur(x)) ∃x (PoodleThat(x) ∧ Dog(x)) ∀x (There(x) → NoPoodleThatDoNotHaveFur(x))	forall x (anythingthat(x) implies doghasfur(x)) exists x (poodlethat(x) and dog(x)) forall x (there(x) implies nopoodlethat-donothavefur(x))	[@every *x [(anythingthat[(?x)] doghasfur[(?x)])] *x [(poodlethat[(?x)] dog[(?x)])] @every *x [(there[(?x)] nopoodlethat-donothavefur[(?x)])]]	forall (anythingthat(x) -: doghasfur(x)) (poodlethat(x) , dog(x)) forall (there(x) -: nopoodlethat-donothavefur(x))	-(+A0-+D0)+(+P1++D1)-(+T0-+N0)	all:x (anythingthat(x) :- doghasfur(x)) x (poodlethat(x) & dog(x)) all:x (there(x) :- nopoodlethat-donothavefur(x))
4480e5d5-495a-4928-a420-a3c74b9268a9	∀x (SingleMammal(x) → Animal(x)) ∀x (Feline(x) → Animal(x)) ∀x (ThiMakeItTrueThatEveryFeline(x) → Mammal(x))	forall x (singlemammal(x) implies animal(x)) forall x (feline(x) implies animal(x)) forall x (thimakeittruethatevereryfeline(x) implies mammal(x))	[@every *x [(singlemammal[(?x)] animal[(?x)])] @every *x [(feline[(?x)] animal[(?x)])] @every *x [(thimakeittruethat-everyfeline[(?x)] mammal[(?x)])]]	forall (singlemammal(x) -: animal(x)) forall (feline(x) -: animal(x)) forall (thimakeittruethat-everyfeline(x) -: mammal(x))	-(+S0-+A0)-(+F0-+A0)-(+T0-+M0)	all:x (singlemammal(x) :- animal(x)) all:x (feline(x) :- animal(x)) all:x (thimakeittruethat-everyfeline(x) :- mammal(x))

Table 7: CLGC transformation example on the working dataset.

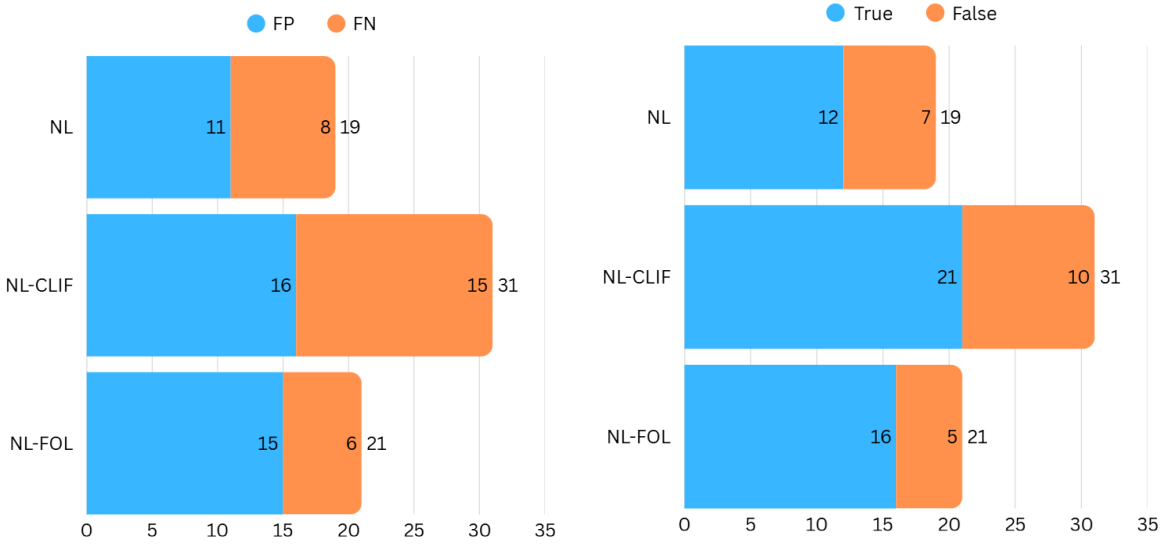

Figure 2: Left: Error analysis of validity prediction of the best SEMEVAL Flan-T5-large model notations. FP = False Positives, FN = False Negatives. Right: Plausibility ground truth of prediction errors for the best SEMEVAL Flan-T5-large models.

Set	Split	Size	Ca	Hy	Di	Co
Working	Train	561	**547**	14	0	0
Working	Val	375	**368**	7	0	0
Working	Test	104	**101**	3	0	0
Evaluation	Eval	191	**189**	2	0	0

Table 8: SEF classification results on the working and evaluation sets: majority syllogism class in bold. Ca = Categorical; Hy = Hypothetical; Di = Disjunctive; Co = Complex.

Dataset	V	IV	P	IP
Pilot	40	40	40	40
Training	480	480	474	486
Evaluation	96	95	95	96

Table 9: Task dataset validity and plausibility intra-distribution. V = Valid; IV = Invalid; P = Plausible; IP = Implausible.

Dataset	V-P	V-I	I-P	I-I
Pilot	20	20	20	20
Training	240	240	234	246
Evaluation	48	48	47	48

Table 10: Task dataset validity and plausibility inter-distribution. V-P = Valid-Plausible; V-I = Valid-Implausible; I-P = Invalid-Plausible; I-I = Invalid-Implausible.

ID	Syllogism	Validity	Plausibility	NL	NL-CLIF	NL-FOL
e773bd8c-fa53-4e9c-8ec6-7d978e0601ac	Every single object can fly. It is known that some boats cannot fly. It follows that some boats are not objects.	true	false		X	
eae77932-d7db-4ce8-b5b1-6c3c951ef553	It is the case that some pencils are white. There are some sheets of paper that are not white. This implies that some sheets of paper are not pencils.	false	true		X	X
fc1a0164-31a4-48aa-ab9d-176a66b93bfe	Anything that is a square is also a quadrilateral. No circle is a square. It follows that no circle is a quadrilateral.	false	true	X		
6725d344-7c13-4d44-a97d-a4a1b89f858d	The category of celestial bodies and the category of planets are mutually exclusive. Not a single planet is a star. It must be the case that a portion of stars are not celestial bodies.	false	false	X	X	
fb637f9c-1c26-4302-9c01-94c061bd352c	The category of chairs and the category of living things do not overlap. The group of living things and the group of inanimate objects are mutually exclusive. A portion of inanimate objects are not chairs.	false	true	X	X	X
fc53be1a-0b75-4995-bcb2-bc3a878a0cb2	"No animal that is a house pet is a feline. Not a single cat is a house pet. This demonstrates that some cats are not felines.	false	false	X		X

Table 11: SEMEVAL sample qualitative error analysis by model: X indicates the model made a mistake in the prediction.

ID	Syllogism	Validity	Plausibility	NL	NL-CLIF	NL-FOL
6725d344-7c13-4d44-a97d-a4a1b89f858d	The category of celestial bodies and the category of planets are mutually exclusive. Not a single planet is a star. It must be the case that a portion of stars are not celestial bodies.	false	false	X	X	X
45d4df27-8269-4ecb-9ead-f6571561f3d5	There is at least one spoon that is a kitchen tool. Some of the utensils are spoons. Consequently, some of the utensils are kitchen tools.	false	true	X	X	X
fc1a0164-31a4-48aa-ab9d-176a66b93bfe	Anything that is a square is also a quadrilateral. No circle is a square. It follows that no circle is a quadrilateral.	false	true	X		X
fb637f9c-1c26-4302-9c01-94c061bd352c	The category of chairs and the category of living things do not overlap. The group of living things and the group of inanimate objects are mutually exclusive. A portion of inanimate objects are not chairs.	false	true		X	X
e773bd8c-fa53-4e9c-8ec6-7d978e0601ac	Every single object can fly. It is known that some boats cannot fly. It follows that some boats are not objects.	true	false		X	
c62f852d-16fd-4eda-b380-b85d9a17f9e2	Every single creature is a mammal. It is true that no mammal is an amphibian. Therefore, it's the case that no amphibian is a creature.	true	false			X

Table 12: FOLIO-SEMEVAL sample qualitative error analysis by model: X indicates the model made a mistake in the prediction.

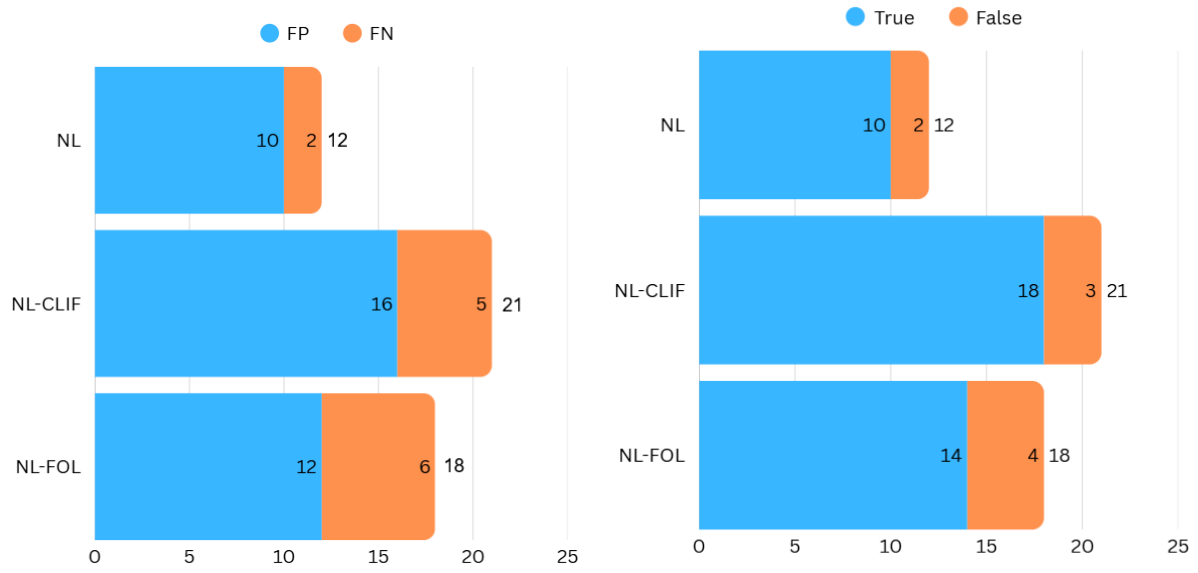

Figure 3: Left: Error analysis of validity prediction of the best FOLIO-SEMEVAL Flan-T5-large model notations. FP = False Positives, FN = False Negatives. Right: Plausibility ground truth of prediction errors for the best FOLIO-SEMEVAL Flan-T5-large models.

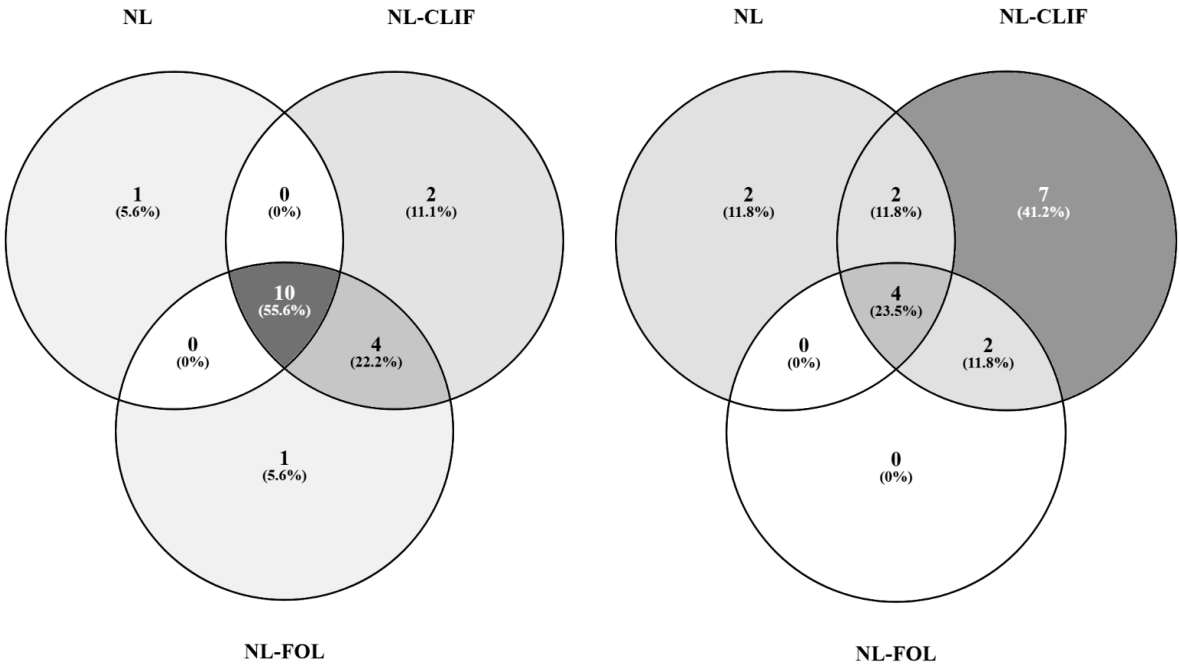

Figure 4: Left: False Positive common errors of the best SEMEVAL Flan-T5-large model notations. Right: False Negative common errors of the best SEMEVAL Flan-T5-large models.

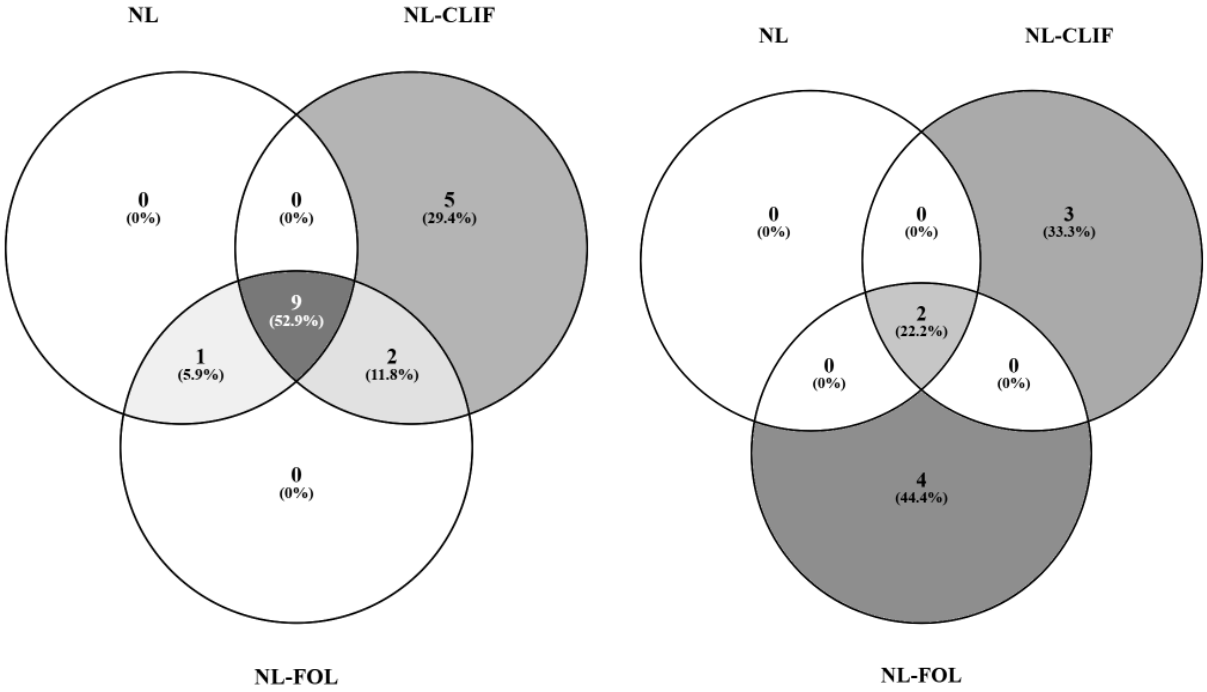

Figure 5: Left: False Positive common errors of the best FOLIO-SEMEVAL Flan-T5-large model notations. Right: False Negative common errors of the best FOLIO-SEMEVAL Flan-T5-large models.