Neuro-Symbolic Logical Reasoning with Textual Entailment

Zacchary Sadeddine and Fabian M. Suchanek

Télécom Paris, Institut Polytechnique de Paris

Abstract. Large Language Models can use logical deduction to answer natural language questions. However, they remain black-boxes and their chains-of-thought can be erroneous. In this paper, we propose to adapt VANESSA, a method for chain-of-thought verification, to the task of reasoning-based question answering. The result is a logic-based, transparent method for answering natural language questions. VANESSA can deliver a formal proof of correctness of its answer using a logical reasoner. Our experiments across various datasets show that the symbolic variant yields high precision, but suffers from low recall due to phrasing differences. The neuro-symbolic variant of VANESSA, which incorporates textual entailment via a black-box model, however, is competitive with the state of the art. We present a demo interface that allows users to interact with the system.

1 Introduction

Reasoning-based question answering is the task of answering a question about a paragraph of text by help of logical deduction. We focus more specifically on yes/no questions, as in the following example:

> Context: If someone likes bread, then they like chocolate or cheese. Anyone who likes eating tomatoes hates even the idea of cheese. Lisa is the biggest tomato lover I know, but she also is a fan of bread.
> Question: Is Lisa fond of chocolate?

This task is useful not only for answering complex questions, but also for gauging whether a system goes beyond a surface-level comprehension of the text. Large Language Models (LLMs) are relatively good at such reasoning tasks, in particular with chain-of-thought prompting. However, LLMs remain black-boxes: chains-of-thought can contain erroneous reasoning steps [4, 9] even when the final answer is correct.

VANESSA is a method to verify the reasoning steps of LLMs [9]. It can run in a symbolic mode (yielding completely transparent verifications), and in a neuro-symbolic mode (which uses natural language inference to bridge variations in phrasing). For example, "is a fan of bread" entails "likes bread", and so the neuro-symbolic VANESSA can use both phrases in a proof.

In this paper, we adapt VANESSA to perform reasoning-based question answering directly, minimizing the use of LLMs in the inference process. Our experiments on multiple logical QA benchmarks compare VANESSA to black-box and neuro-symbolic competitors, and show that (1) the symbolic variant of VANESSA consistently achieves the highest precision across all datasets and competitors when it finds a proof and (2) the neuro-symbolic variant of VANESSA is competitive in overall performance with purely neural methods, while additionally delivering a formal proof tree. Our graphical user interface allows the audience to interact with the system, pose questions, and inspect the answers and proof trees.

2 Related Work

LLMs have been used extensively for all kinds of reasoning problems. Chain-of-thought prompting [15] has further increased model performance while giving the user access to a proof, but hallucinations and formal errors still can be present [4, 9]. To address these issues, several works have investigated neuro-symbolic methods that use external tools such as calculators or knowledge bases in combination with LLMs, increasing performance on a variety of tasks [14, 2, 3, 12]. For logical reasoning on text, the most common approach has been to make an LLM parse the input into a machine-readable format such as Prolog [6, 1, 16] or First-Order Logic [7], and then perform reasoning over these structures with theorem provers. However, the parsing into a logical formalism is a black-box step: If we don't trust the LLM on formal reasoning in a chain-of-thought, then there is no reason to trust it on the translation to formal logic.

Our work, in contrast, proposes a fully symbolic and transparent reasoning method. To increase recall, our method can be run in a neuro-symbolic variant, which uses an LLM, but only for Natural Language Inference (NLI). Thereby, the area of distrust is reduced to a single atomic task, on which LLMs usually perform well. Besides, an NLI step is usually trivial to verify manually.

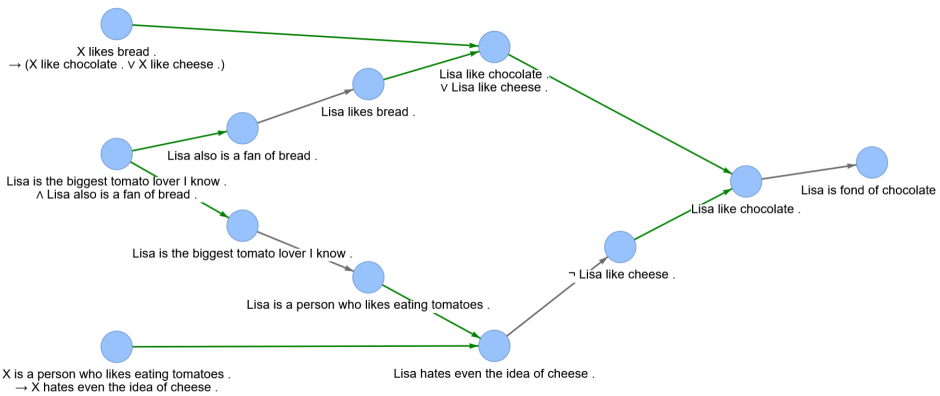
3 VANESSA

VANESSA is a method to verify the reasoning within a chain-of-thought [9]. The input to VANESSA is a context, a boolean question, and a chain-of-thought that is composed of reasoning steps, each consisting of premises and a conclusion. VANESSA then checks every single reasoning step and outputs "Correct" if every step is valid and every premise is grounded in the context or in previous conclusions. VANESSA operates in three phases: (1) a shallow symbolic parsing of the context and the question, (2) an augmentation of the logical forms through Natural Language Inference (NLI), and (3) symbolic reasoning. Step (2) can be performed symbolically by string matching (resulting in the fully symbolic variant of VANESSA) or with an LLM (yielding a neuro-symbolic variant, which is more robust to variations in phrasing).

The present work adapts VANESSA to perform question answering directly, without requiring a given chain-of-thought. The input to our adapted method is a context, consisting of rules and facts in natural language, and a boolean question (as in the example in the introduction). The context has to be self-contained, i.e., no external

Figure 1: Experimental results (left) and proof tree shown by our demo for our running example (right)

	Accuracy	Prec	Rec	F1
ProofWriter				
⋆⋆ VANESSA symb.	88.27	97.56	83.33	89.89
VANESSA neuro	65.90	63.03	78.12	69.77
LINC	73.55	94.52	71.88	81.66
CoT	45.29	40.19	44.79	42.37
Direct	27.04	24.83	38.54	30.2
ProntoQA				
VANESSA symb.	39.01	96.77	23.62	37.97
⋆⋆ VANESSA neuro	84.18	85.03	98.43	91.24
LINC	65.87	85.09	76.38	80.5
CoT	68.92	74.48	85.04	79.41
Direct	59.16	59.38	74.8	66.2
FOLIO				
VANESSA symb.	36.05	100.0	2.96	5.75
⋆ VANESSA neuro	49.52	59.14	40.74	48.25
LINC	34.12	86.84	24.44	38.14
CoT	55.77	57.06	74.81	64.74
Direct	55.92	53.3	71.85	61.2
LogicBench BD				
VANESSA symb.	50.00	x	x	x
⋆ VANESSA neuro	52.28	55.56	50.0	52.63
LINC	24.91	40.0	20.0	26.67
CoT	59.12	57.69	75.0	65.22
Direct	63.69	62.5	75.0	68.18



The table shows accuracy, as well as micro-averaged precision, recall, and F1 for the "yes" & "no" classes. The background color indicates white-box, gray-box, and black-box approaches. The star ⋆ is for the best white-box/gray-box approach; the double star ⋆⋆ for a white-box/gray-box approach that beats even black-box approaches.

knowledge is needed to answer the question. There are three possible answers: "yes", "no" and "unknown" (if the context does not permit a definite conclusion).

We transform this input into a pseudo-reasoning step, which has the entire context as premises, and the question (in the form of an affirmative sentence) as the conclusion. Our adapted VANESSA then tries to validate the reasoning step. If this succeeds, the answer to the question is "yes". If it fails, our method tries to validate the negation of the conclusion. If that succeeds, the answer is "no". Otherwise the answer is "unknown". When VANESSA successfully finds an answer, it automatically constructs a proof tree, which is presented as an explanation supporting the answer.

4 Experiments

We evaluate our method on several logical reasoning datasets: ProofWriter [13] ("Depth 5, Open World Assumption" Dev set), ProntoQA [10] (using the 100 first instances of the 4-hop Composed Random set from ProntoQA-OOD [11]), FOLIO [5], and LogicBench [8] (for which we manually relabeled negative ground truth examples as either "no" or "unknown" and subsampled the datasets to achieve a balance between the possible answers). We compare several approaches: black-box methods (prompt-instructed LLM for direct answer or for chain-of-thought), gray-box neuro-symbolic approaches (LINC [7] and VANESSA neuro-symbolic) and white-box (symbolic VANESSA). Experiments use LLaMa3-8B-Instruct as the model, with task-specific prompts.

Figure 1 (left) shows that the white-box symbolic VANESSA performs as expected: Whenever it delivers results, these consistently have the highest precision. It even reaches best overall accuracy on ProofWriter. However, the method falls behind on recall because of its inability to deal with phrasing variations. On the LogicBench datasets, this issue goes so far that the method fails to deliver any verdict at all, always outputting "unknown", which results in an accuracy of 50% on subsets where half the ground truth labels are "unknown".

Among the gray-box approaches, neuro-symbolic VANESSA consistently achieves higher accuracy than LINC. On several datasets, VANESSA beats even the black-box approaches, a feat that LINC does not achieve. As expected, among the black-box approaches, chain-of-thought prompting generally outperforms direct prompting approach, albeit only on 3 out of the 5 datasets.

Overall, our experiments thus show that symbolic and neuro-symbolic methods can compete with black-box models in terms of accuracy, with the neuro-symbolic VANESSA emerging as the best-performing gray-box method.

5 Demo

A demonstration of our system is available at https://vanessa-demo.org/, or can be downloaded for local use. The user can input premises and a conclusion for a logical reasoning problem, and run VANESSA in either symbolic or neuro-symbolic mode. The neuro-symbolic variant uses LLaMa3.2-3B. It is somewhat slow in the online interface due to computational requirements, but can be faster when run locally.

When VANESSA finds a solution to the reasoning problem, the interface displays a proof tree (Figure 1 right). Gray arrows indicate textual entailment, while green arrows indicate logical deduction. The interface also shows the parsed input sentences, the detected entailments and a linearized textual proof, allowing users to trace the reasoning process in a transparent way.

In the demo, users can play around with the preset examples that the GUI offers from several benchmarks. Users can also modify the examples, for example by rephrasing sentences to test the system's robustness, adding negations or changing conclusions. Finally, they can also submit their own reasoning problems and see if the system can give the correct response.

6 Conclusion

We have presented an adaptation of the VANESSA method for answering natural language questions. We hope that this work paves the way for the development of more explainable and transparent logical reasoning systems. All code and data is available at https://github.com/dig-team/VANESSA/tree/demo.

Acknowledgements

This work was partially funded by the NoRDF project (ANR-20-CHIA-0012-01).

References

[1] N. Borazjanizadeh and S. T. Piantadosi. Reliable reasoning beyond natural language. *arXiv preprint arXiv:2407.11373*, 2024.

[2] M. Fang, S. Deng, Y. Zhang, Z. Shi, L. Chen, M. Pechenizkiy, and J. Wang. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17985–17993, 2024.

[3] Y. Ge, S. Romeo, J. Cai, R. Shu, M. Sunkara, Y. Benajiba, and Y. Zhang. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. *arXiv preprint arXiv:2502.01630*, 2025.

[4] O. Golovneva, M. P. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, and A. Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=xYlJRpzZtsY.

[5] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, L. Benson, L. Sun, E. Zubova, Y. Qiao, M. Burtell, D. Peng, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, S. Joty, A. R. Fabbri, W. Kryscinski, X. V. Lin, C. Xiong, and D. Radev. Folio: Natural language reasoning with first-order logic, 2022.

[6] J. Lee and W. Hwang. Symba: Symbolic backward chaining for multi-step natural language reasoning. *arXiv preprint arXiv:2402.12806*, 2024.

[7] T. Olausson, A. Gu, B. Lipkin, C. Zhang, A. Solar-Lezama, J. Tenenbaum, and R. Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.313. URL https://aclanthology.org/2023.emnlp-main.313.

[8] M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, and C. Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL https://aclanthology.org/2024.acl-long.739.

[9] Z. Sadeddine and F. M. Suchanek. Verifying the steps of deductive reasoning chains. In *Proceedings of the 2025 Conference of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Findings)*. Association for Computational Linguistics, 2025.

[10] A. Saparov and H. He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qFVVBzXxR2V.

[11] A. Saparov, R. Y. Pang, V. Padmakumar, N. Joshi, M. Kazemi, N. Kim, and H. He. Testing the general deductive reasoning capacity of large language models using OOD examples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=MCVfX7HgPO.

[12] F. M. Suchanek and A. T. Luu. Knowledge Bases and Language Models: Complementing Forces. In *RuleML+RR invited paper*, 2023.

[13] O. Tafjord, B. Dalvi, and P. Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317.

[14] H. Wang, H. Xin, C. Zheng, L. Li, Z. Liu, Q. Cao, Y. Huang, J. Xiong, H. Shi, E. Xie, J. Yin, Z. Li, H. Liao, and X. Liang. Lego-prover: Neural theorem proving with growing libraries, 2023.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

[16] S. Yang, X. Li, L. Cui, L. Bing, and W. Lam. Neuro-symbolic integration brings causal and reliable reasoning proofs. *arXiv preprint arXiv:2311.09802*, 2023.