Neuro-Symbolic Logical Reasoning with Textual Entailment

Zacchary Sadeddine
Télécom Paris
zacchary.sadeddine@telecom-paris.fr

Fabian M. Suchanek
Télécom Paris
fabian@suchanek.name

Abstract

Large Language Models can answer natural language questions with reasoning, but they remain black boxes. In this paper, we propose to adapt VANESSA, a method for chain-of-thought verification, to the task of question answering. The result is a fully symbolic and transparent method for answering natural language questions with logical deduction. VANESSA can deliver a formal proof of correctness of the answer by a logical reasoner. Our experiments across a variety of datasets show that this method yields high precision, but suffers from low recall due to phrasing differences. The neuro-symbolic variant of VANESSA, which allows for textual entailment performed by a black box model, however, is competitive with the state of the art.

1 Introduction

Reasoning-based question answering is the task of answering a yes/no question on a paragraph of text by help of logical deduction. This means that the answer cannot just be found in the text, but has to be deduced in one or more logical reasoning steps from the information given in the text (Figure 1). This task is of interest not just because it allows answering complex questions, but also because it allows gauging whether a QA system goes beyond surface-level comprehension of the text. Large Language Models (LLMs) are relatively good at such reasoning tasks. However, they remain black boxes: If an LLM says “yes” (or “no”), we have no guarantee that this answer is correct. The only way to verify the answer is to compute the answer ourselves and to compare it to the LLM answer – which defies the purpose of using an LLM in the first place. The picture is slightly different with techniques such as Chain-of-Thought: here, the model can show its reasoning explicitly step by step. However, each of these steps might still be incorrect. Interestingly, the steps can be incorrect even if the answer is correct, thus making the answer right for the wrong reasons. Indeed, our previous work[1] shows that LLMs frequently produce hallucinations, omissions, or errors in their reasoning steps.

Our previous work has shown how these reasoning steps can be verified in a symbolic manner, by a method we called VANESSA. For each reasoning step in the Chain-of-Thought, VANESSA parses sentences into a syntax tree, applies a set of symbolic tree transformations, and verifies the conclusion by help of a symbolic logical reasoner. In a neuro-symbolic variant, VANESSA uses Natural Language Inference (NLI) to bridge differences in phrasing.

In this paper, we show that VANESSA can be adapted to answer the initial question right away, instead of verifying the reasoning steps produced by an LLM. For this purpose, VANESSA has to be adapted to not only prove if the conclusion is correct, but also to prove that it is incorrect (in which case the output answer will be “no”). Such answers are accompanied by a set of transparent steps that lead from the input text to the answer, and come with a formal proof of correctness. Figure 1 shows an example. Note how the neuro-symbolic variant of VANESSA can understand that “Lisa likes chocolate” entails “Lisa is fond of chocolate”, and thereby answer the question correctly.

Our experiments on several logical QA benchmarks and with black box competitors as well as neuro-symbolic competitors show that (1) the symbolic VANESSA consistently has the highest precision across all datasets and competitors when it finds a proof and (2) in its neuro-symbolic variant, VANESSA is competitive in overall performance with purely neural methods, while at the same time delivering a formal proof tree.

[1] This previous work has been submitted to the ACL Rolling Review December 2024 Cycle, where it has received a meta-review score of 4/5 and will be committed to the ACL conference. For transparency, we have added that paper to https://anonymous.4open.science/r/LLMVerifier-403B.

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?*

LLM

yes

LLM with Chain-of-Thought:

Let's think step by step: Lisa is the biggest tomato lover I know, but she also is a fan of bread.
Someone who likes bread will like either chocolate or cheese.
Therefore, Lisa likes chocolate or cheese.
Since she loves tomatoes, she hates the very idea of cheese. Therefore, Lisa likes chocolate

VANESSA symbolic:

1. Pattern "Everyone..." applied to first sentence: $\forall x : x\_likes\_bread \Rightarrow (x\_like\_chocolate \lor x\_like\_cheese)$
2. 1 instantiated with Lisa: $Lisa\_likes\_bread \Rightarrow (Lisa\_like\_chocolate \lor Lisa\_like\_cheese)$
3. Translation of third sentence: $Lisa\_is\_the\_biggest\_tomato\_lover\_I\_know \lor Lisa\_is\_a\_fan\_of\_bread$

Cannot deduce $Lisa\_is\_fond\_of\_chocolate$

VANESSA neuro-symbolic:

1. Pattern "Everyone..." applied to first sentence: $\forall x : x\_likes\_bread \Rightarrow (x\_like\_chocolate \lor x\_like\_cheese)$
2. 1 instantiated with Lisa: $Lisa\_likes\_bread \Rightarrow (Lisa\_like\_chocolate \lor Lisa\_like\_cheese)$
3. Translation of third sentence: $Lisa\_is\_the\_biggest\_tomato\_lover\_I\_know \lor Lisa\_is\_a\_fan\_of\_bread$
4. Textual entailment $Lisa\_is\_a\_fan\_of\_bread \rhd Lisa\_likes\_bread$
5. Modus Ponens from 3 and 4: $Lisa\_like\_chocolate \lor Lisa\_like\_cheese$
6. Textual entailment $Lisa\_is\_the\_biggest\_tomato\_lover\_I\_know \rhd Lisa\_likes\_eating\_tomatoes$
7. Translation and instantiation of second sentence: $Lisa\_likes\_eating\_tomatoes \Rightarrow Lisa\_hates\_even\_the\_idea\_of\_cheese$
8. Modus Ponens from 6 and 7: $Lisa\_hates\_even\_the\_idea\_of\_cheese$
9. Or-elimination from 2 and 4: $Lisa\_likes\_chocolate$
10. Textual Entailment $Lisa\_likes\_chocolate \rhd Lisa\_is\_fond\_of\_chocolate$

yes

Figure 1: Illustration of reasoning-based question answering (fictional example)

2 Related Work

LLMs have been used extensively for all kinds of reasoning problems. Chain-of-Thought prompting (Wei et al., 2022) has further increased the model performance, while giving the user access to a proof. However, the whole reasoning still relies on the LLM, which can hallucinate and make formal errors. For this reason, 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 (Wang et al., 2023; Fang et al., 2024; Ge et al., 2025). 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 (Lee and Hwang, 2024; Borazjanizadeh and Piantadosi, 2024; Yang et al., 2023) or First-Order Logic (Olausson et al., 2023), and perform reasoning on 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, there is no reason to trust it on the translation to formal logic.

Our work proposes a fully symbolic and transparent reasoning instead. To increase recall, our method can be combined with 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 extremely well. In addition, NLI is usually trivial to verify manually; in Figure 1 it amounts to checking if "Lisa likes chocolate" contradicts "Lisa is fond of chocolate".

3 VANESSA

VANESSA was introduced in our prior work as a method to verify the reasoning in a chain-of-thought of a language model. The input to VANESSA is a context, a boolean question, and a chain-of-thought that consists of reasoning steps. Each reasoning step consists of premises and a conclusion. VANESSA then checks every single reasoning step and outputs "Correct" iff every step is correct and every premise is grounded in the context or in previous conclusions.

In the present work, we adapt VANESSA to answer the question right away, without a given chain-of-thought. The input to our method is a context, consisting of rules and facts in natural language, and a boolean question. The question is to be answered under the Open World Assumption, meaning there are three possible answers: “Yes”, “No” and “Unknown” (where the latter is given when the context does not allow drawing any conclusion about the question with certainty). 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. Then our adapted VANESSA 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 validates a reasoning step successfully, it automatically constructs a proof tree, which we can show as an explanation of the answer.

We now briefly recap how VANESSA validates a reasoning step. The method proceeds in three phases: a shallow parsing of the context and the question (explained in Section 3.1); an augmentation of the logical forms through NLI (used exclusively in the neuro-symbolic variant of VANESSA, see Section 3.2); and symbolic reasoning (Section 3.3).

3.1 Logic Transformation.

The purpose of the first phase is to transform every sentence from the input into a logical form, with operators linking independent and grammatically correct atomic sentences. We perform co-reference resolution (with LingMess Otmazgin et al., 2023) on the whole input, and then constituency parsing (Kitaev et al., 2019) on each sentence. Each tree is then recursively transformed using tree regular expressions (adapted from Graphene (Niklaus et al., 2016)). Some patterns specifically aim at detecting universal quantification in the sentence.This strategy allows us to ensure that every resulting tree is a well-formed sentence, and keeps the syntax and semantics of the sentence intact. For example, “Alex plays football and eats pasta” is transformed into `Alex plays football` $\wedge$ `Alex eats pasta`.

In the end, all universal quantifications are instantiated with definite noun phrases from the premises and conclusion.

3.2 Natural Language Inference

The previous phase has transformed the context into a set of atomic sentences linked by operators. These sentences are replaced by identifiers to feed them into a reasoner. The problem is that the reasoner cannot see semantic relationships between sentences such as “Lisa hates chocolate” (which will become one identifier, say A), and “Lisa loves chocolate” (which will be another identifier, say B). Indeed, the purely symbolic variant of VANESSA can see equivalences between sentences only if they are identical (and hence receive the same identifier). The neuro-symbolic variant of VANESSA, in contrast, applies Natural Language Inference (NLI) on pairs of atomic statements from the premises and the conclusion. If the NLI module outputs an entailment between A and B, we add $A \Rightarrow B$ to our set of formulas (and $A \Rightarrow \neg B$ if it outputs a contradiction). As in Helwe et al. (2022), we write $A \triangleright B$ for A entails B, and $A \blacktriangleright B$ for A contradicts B. Performing NLI on pairs of atomic statements makes the task easier for the model than having to process several sentences at once, and pushes the bigger part of reasoning to the symbolic prover. Testing all combinations of atomic statements would be both computationally expensive and susceptible to larger errors (as NLI is not 100% reliable). For this reason, we designed a strategy for choosing the statements to pair that reduces the number of pairs as much as possible while still covering all interesting cases.

3.3 Natural Deduction

The final phase of VANESSA aims to deduce the conclusion from the formulas given by the context and those generated during the NLI phase. We built a Natural Deduction Solver (Gentzen, 1935) to this end, which performs a bidirectional search (Pollock, 1999). This method allows the search to prioritize “easy” deductions and restrains the search space to elements that will be directly useful to reach the final objective. Natural Deduction allows us to choose the deduction rules we want to apply. Most notably, we can exclude the rule of material implication, which can lead to counter-intuitive reasonings. The reader is invited to find more details in our original work on VANESSA.

4 Experiments

We test our method with several datasets. Each dataset contains instances of the form (C, Q, O, A), where C is the context, Q is a boolean question, O the set of answer options (which is always "True", "False", and "Unknown"), and $A \in O$ is the ground-truth answer. The information needed to answer the question Q is present in the context C, or can be derived through deductive reasoning to arrive at the correct answer A.

4.1 Datasets

We evaluate our method on several logical reasoning-based question answering datasets: ProofWriter, FOLIO, ProntoQA and LogicBench.

ProofWriter (Tafjord et al., 2021) is a question answering dataset that contains proofs with intermediate steps. This dataset was generated synthetically using small ontologies, and hence contains short and simple sentences with a limited vocabulary. We used 165 question instances from the "Depth 5, Open World Assumption" DEV set.

ProntoQA (Saparov and He, 2023) contains proofs with intermediate reasoning steps for each question. It was also generated using hierarchical ontologies, but uses more diverse and complex reasoning patterns than ProofWriter. We generated positive and negative QA instances using the 100 first instances of the 4-hop Composed Random set from ProntoQA-OOD (Saparov et al., 2023), which has the particularity of using fictional words (e.g. "zumpuses").

FOLIO (Han et al., 2022) is a reasoning-based question answering dataset containing a wide array of problems and reasoning patterns. While the previous two datasets are restricted in their syntax, FOLIO contains sentences with a large variation in formulations, words, and entities. It is based on real-life instances and examples.

LogicBench (Parmar et al., 2024) is a deductive question answering dataset that systematically covers a large array of reasoning patterns. It contains single-step reasoning problems, which have been rephrased into natural language by an LLM, ensuring a large syntactic diversity. We used the propositional logic Hypothetical Syllogism, Disjunctive Syllogism, Constructive Dilemma, Destructive Dilemma, and Bidirectional Dilemma subsets. Contraposition and Material Implication were not considered because they are concerned with formal logic rather than natural language reasoning. The Modus Tollens subset was excluded because manual inspection showed that the reformulation by the LLM lead to contradictions between the context and the answer. The original dataset does not consider the Open World Assumption, and thus makes no distinction between "False" and "Unknown". Hence we manually relabeled negative ground truths as either "False" or "Uncertain", and subsampled the datasets to achieve a balance between the possible answers.

4.2 Competitors

We compare our approach with several other neural and neuro-symbolic methods.

Direct LLM. We ask a language model directly for the answer to the question given the context. This method is a black-box method. We used two models that were few-shot prompted for the task: Ministral (8B) and LLaMa3-8B-Instruct.

CoT LLM. We ask the same language models to answer the question by a chain-of-thought. This method is more transparent, as it gives a proof. However, there is no guarantee that this chain is correct, which still makes this method black-box.

LINC. The LINC framework (Olausson et al., 2023) transforms each instance to first-order logic by help of an LLM, and then verifies the conclusion using a formal theorem prover (Olausson et al., 2023; Pan et al., 2023). This method is more transparent than a direct LLM answer, although the transformation to first order logic is still opaque. We consider this method gray-box. The original paper uses GPT 3.5-turbo in its experiments, which we replaced with the same Open Source models as before, making the results easier to reproduce.

VANESSA neuro-symbolic. This method is completely transparent up to the natural language inference, which makes it grey-box. We used once again the same LLaMa3 and Ministral models.

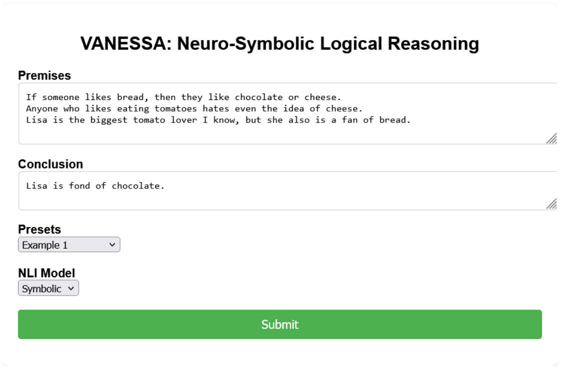
VANESSA symbolic. This method is completely transparent.

4.3 Results

Table 1 shows the performance of the different approaches with LLaMa 3 on all datasets (the results with Ministral are in Appendix A and do not differ much). The symbolic VANESSA performs as expected: Whenever it delivers results, these consistently have the highest precision, notably reaching best performance on ProofWriter. However, the method falls behind on recall because of its inability to deal with phrase variations. On the LogicBench datasets, this problem goes so far that the method is unable to deliver a verdict at all, and always says "unknown" – which gives an accuracy of 50% on subsets where half the ground truth labels are “unknown”. If we look at the gray-box approaches, the neuro-symbolic VANESSA always has a better accuracy than LINC, with only one exception. On half of the datasets, VANESSA beats even the black-box approaches – a feat that LINC does not achieve. Among these pproaches, the CoT approach has a better accuracy that the Direct approach, as one might expect, albeit only on 5 out of the 8 datasets.

Our experiments thus show that symbolic and neuro-symbolic methods can compete with black-box models in terms of accuracy, and that the neuro-symbolic VANESSA is generally the best-performing gray-box model.

5 Demonstration



VANESSA: Neuro-Symbolic Logical Reasoning

Premises

```
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.
```

Conclusion

```
Lisa is fond of chocolate.
```

Presets: Example 1

NLI Model: Symbolic

Figure 2: Welcome page of our Web interface

A demonstration of our system is available at https://vanessa.r2.enst.fr. Figure 2 shows the home page, which invites the user to input premises and a conclusion for a logical reasoning problem. The user can then choose to run VANESSA in the symbolic mode, or in the neuro-symbolic mode. The latter relies on models that require at least 20GB of VRAM to run, and so we cannot deploy them in the online interface. Hence, the neuro-symbolic mode of VANESSA is available only in the hands-on version of our demo.

When VANESSA finds a solution to the reasoning problem, the interface shows the Natural Deduction proof in the form of an interactive directed acyclic graph (Figure 3). This graph shows which statements entail which other statements (in gray arrows), and which statements contribute to the logical deduction of which other statements (in green arrows). We also show the parsed input sentences, the detected entailments, and the linearized proof in textual form (Figure 4). This allows users to

	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 HS				
VANESSA symb.	50.00	x	x	x
VANESSA neuro	46.42	40.0	15.0	21.82
⋆ LINC	54.77	61.54	40.0	48.49
CoT	58.35	55.56	87.5	67.96
Direct	50.00	x	x	
LogicBench DS				
VANESSA symb.	0.00	x	x	x
⋆ VANESSA neuro	54.56	88.0	55.0	67.69
LINC	29.47	84.62	27.5	41.51
CoT	65.97	87.1	67.5	76.06
Direct	31.75	38.71	30.0	33.8
LogicBench CD				
VANESSA symb.	50.00	x	x	x
⋆⋆ VANESSA neuro	56.84	56.0	70.0	62.22
LINC	52.28	75.0	45.0	56.25
CoT	52.28	51.35	95.0	66.67
Direct	50.00	48.65	90.0	63.16
LogicBench DD				
VANESSA symb.	50.00	x	x	x
⋆⋆ VANESSA neuro	52.28	50.0	45.0	47.37
LINC	31.75	31.25	25.0	27.78
CoT	50.00	47.83	55.0	51.17
Direct	50.00	50.0	70.0	58.33
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

Table 1: Accuracy, as well as micro-averaged precision, recall, and F1 for the positive and negative classes. ⋆ for best whitebox/graybox approach. ⋆⋆ for whitebox/graybox approach that beats even blackbox approaches. All approaches run with Llama3.

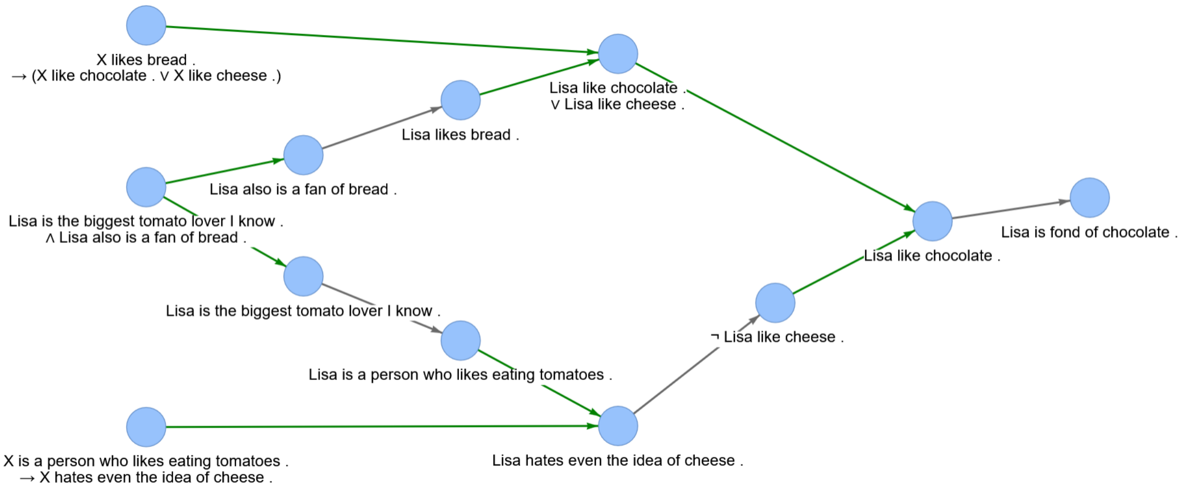
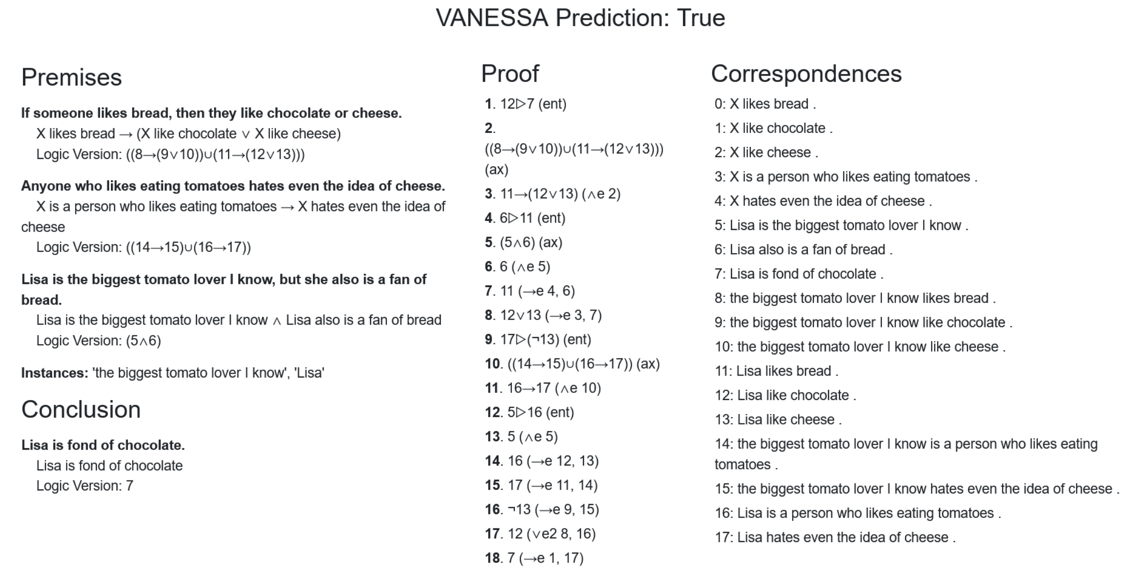

X likes bread .
→ (X like chocolate . ∨ X like cheese .)
Lisa like chocolate .
∨ Lisa like cheese .
Lisa likes bread .
Lisa also is a fan of bread .
Lisa is the biggest tomato lover I know .
∧ Lisa also is a fan of bread .
Lisa is fond of chocolate .
Lisa like chocolate .
Lisa is the biggest tomato lover I know .
¬ Lisa like cheese .
Lisa is a person who likes eating tomatoes .
X is a person who likes eating tomatoes .
→ X hates even the idea of cheese .
Lisa hates even the idea of cheese .

Figure 3: Proof graph for our running example



VANESSA Prediction: True

Premises

If someone likes bread, then they like chocolate or cheese.
X likes bread → (X like chocolate ∨ X like cheese)
Logic Version: ((8→(9∨10))∪(11→(12∨13)))

Anyone who likes eating tomatoes hates even the idea of cheese.
X is a person who likes eating tomatoes → X hates even the idea of cheese
Logic Version: ((14→15)∪(16→17))

Lisa is the biggest tomato lover I know, but she also is a fan of bread.
Lisa is the biggest tomato lover I know ∧ Lisa also is a fan of bread
Logic Version: (5∧6)

Instances: 'the biggest tomato lover I know', 'Lisa'

Conclusion

Lisa is fond of chocolate.
Lisa is fond of chocolate
Logic Version: 7

Proof

1. 12⊳7 (ent)
2. ((8→(9∨10))∪(11→(12∨13))) (ax)
3. 11→(12∨13) (∧e 2)
4. 6⊳11 (ent)
5. (5∧6) (ax)
6. 6 (∧e 5)
7. 11 (→e 4, 6)
8. 12∨13 (→e 3, 7)
9. 17⊳(¬13) (ent)
10. ((14→15)∪(16→17)) (ax)
11. 16→17 (∧e 10)
12. 5⊳16 (ent)
13. 5 (∧e 5)
14. 16 (→e 12, 13)
15. 17 (→e 11, 14)
16. ¬13 (→e 9, 15)
17. 12 (∨e2 8, 16)
18. 7 (→e 1, 17)

Correspondences

0: X likes bread .
1: X like chocolate .
2: X like cheese .
3: X is a person who likes eating tomatoes .
4: X hates even the idea of cheese .
5: Lisa is the biggest tomato lover I know .
6: Lisa also is a fan of bread .
7: Lisa is fond of chocolate .
8: the biggest tomato lover I know likes bread .
9: the biggest tomato lover I know like chocolate .
10: the biggest tomato lover I know like cheese .
11: Lisa likes bread .
12: Lisa like chocolate .
13: Lisa like cheese .
14: the biggest tomato lover I know is a person who likes eating tomatoes .
15: the biggest tomato lover I know hates even the idea of cheese .
16: Lisa is a person who likes eating tomatoes .
17: Lisa hates even the idea of cheese .

Figure 4: Linearization and output details

trace the reasoning steps in a transparent way. In case of an error, users can see whether the error comes from the parsing or the NLI.

During the demo, users can start by playing around with some of the preset examples that the GUI offers from several benchmarks. Users can then modify the examples, for example by changing the phrasing (to see if the system is robust), or by adding negations or different 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 that can be used to answer natural language questions – either by fully symbolic, transparent reasoning, or by neuro-symbolic reasoning. Our experiments on a variety of benchmarks show that the symbolic variant can achieve a high precision at a somewhat reduced recall. The neuro-symbolic variant, however, performs on par with black-box models or even better, while still being more transparent. In a hands-on demo, users can play with the system, submit their own logical riddles, and try to trick our system. We hope that our work paves the way for the development of more explainable and more transparent logic reasoning models.

All our code and data is available for review at `https://anonymous.4open.science/r/LLMVerifier-403B`, and the video at `https://youtu.be/eWSop2Mayow`.

References

Nasim Borazjanizadeh and Steven T Piantadosi. 2024. Reliable reasoning beyond natural language. *arXiv preprint arXiv:2407.11373*.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17985–17993.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Monica Sunkara, Yassine Benajiba, and Yi Zhang. 2025. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. *arXiv preprint arXiv:2502.01630*.

Gerhard Gentzen. 1935. Untersuchungen Über das Logische Schließen. I. *Mathematische Zeitschrift*, 35:176–210.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic.

Chadi Helwe, Simon Coumes, Chloé Clavel, and Fabian M. Suchanek. 2022. TINA: Textual Inference with Negation Augmentation. In *EMNLP Find.*

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Jinu Lee and Wonseok Hwang. 2024. Symba: Symbolic backward chaining for multi-step natural language reasoning. *arXiv preprint arXiv:2402.12806*.

Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas. 2016. A sentence simplification system for improving relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 170–174, Osaka, Japan. The COLING 2016 Organizing Committee.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

John Pollock. 1999. Natural deduction. *an unpublished manuscript is available at http://johnpollock.us/ftp/OSCAR-web-page/PAPERS/Natural-Deduction. pdf.*

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Lin Li, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, Heng Liao, and Xiaodan Liang. 2023. Lego-prover: Neural theorem proving with growing libraries.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Sen Yang, Xin Li, Leyang Cui, Lidong Bing, and Wai Lam. 2023. Neuro-symbolic integration brings causal and reliable reasoning proofs. *arXiv preprint arXiv:2311.09802*.

A Full Results

Dataset	**Method**	**Model**	**Accuracy**	Prec	Rec	F1
ProofWriter	VANESSA	Symbolic	**88.27**	97.56	83.33	89.89
		LLaMa3	65.90	63.03	78.12	69.77
		Ministral	68.25	66.02	70.83	68.34
	LINC	LLaMa3	73.55	94.52	71.88	81.66
		Ministral	82.38	82.0	85.42	83.68
	CoT	LLaMa3	45.29	40.19	44.79	42.37
		Ministral	50.59	46.55	56.25	50.94
	Direct	LLaMa3	27.04	24.83	38.54	30.2
		Ministral	32.34	31.93	55.21	40.46
ProntoQA	VANESSA	Symbolic	39.01	96.77	23.62	37.97
		LLaMa3	84.18	85.03	98.43	91.24
		Ministral	**95.78**	97.64	97.64	97.64
	LINC	LLaMa3	65.87	85.09	76.38	80.5
		Ministral	44.51	78.16	53.54	63.55
	CoT	LLaMa3	68.92	74.48	85.04	79.41
		Ministral	70.75	77.94	83.46	80.61
	Direct	LLaMa3	59.16	59.38	74.8	66.2
		Ministral	45.12	45.0	56.69	50.17
FOLIO	VANESSA	Symbolic	36.05	100.0	2.96	5.75
		LLaMa3	49.52	59.14	40.74	48.25
		Ministral	47.37	76.60	26.47	39.34
	LINC	LLaMa3	34.12	86.84	24.44	38.14
		Ministral	39.90	89.36	31.11	46.15
	CoT	LLaMa3	55.77	57.06	74.81	64.74
		Ministral	**56.74**	56.65	72.59	63.64
	Direct	LLaMa3	55.92	53.3	71.85	61.2
		Ministral	46.15	45.54	68.15	54.6
LogicBench HS	VANESSA	Symbolic	50.00	x	x	x
		LLaMa3	46.42	40.0	15.0	21.82
		Ministral	45.52	33.33	10.0	15.38
	LINC	LLaMa3	54.77	61.54	40.0	48.49
		Ministral	54.77	80.0	30.0	43.64
	CoT	LLaMa3	58.35	55.56	87.5	67.96
		Ministral	**84.59**	79.17	95.0	86.37
	Direct	LLaMa3	50.00	x	x	x
		Ministral	44.04	43.75	87.5	58.33
LogicBench DS	VANESSA	Symbolic	0.00	x	x	x
		LLaMa3	54.56	88.0	55.0	67.69
		Ministral	47.72	90.48	47.5	62.3
	LINC	LLaMa3	29.47	84.62	27.5	41.51
		Ministral	8.94	100.0	5.0	9.52
	CoT	LLaMa3	65.97	87.1	67.5	76.06
		Ministral	**72.81**	81.08	75.0	77.92
	Direct	LLaMa3	31.75	38.71	30.0	33.8
		Ministral	47.72	47.5	47.5	47.5
LogicBench CD	VANESSA	Symbolic	50.00	x	x	x
		LLaMa3	56.84	56.0	70.0	62.22
		Ministral	50.00	x	x	x
	LINC	LLaMa3	52.28	75.0	45.0	56.25
		Ministral	47.72	70.0	35.0	46.67
	CoT	LLaMa3	52.28	51.35	95.0	66.67
		Ministral	**59.12**	55.88	95.0	70.37
	Direct	LLaMa3	50.00	48.65	90.0	63.16
		Ministral	38.60	37.5	75.0	50.0
LogicBench DD	VANESSA	Symbolic	50.00	x	x	x
		LLaMa3	52.28	50.0	45.0	47.37
		Ministral	**56.84**	63.64	35.0	45.16
	LINC	LLaMa3	31.75	31.25	25.0	27.78
		Ministral	38.60	50.0	35.0	41.18
	CoT	LLaMa3	50.00	47.83	55.0	51.17
		Ministral	52.28	50.0	80.0	61.54
	Direct	LLaMa3	50.00	50.0	70.0	58.33
		Ministral	0.00	x	x	x
LogicBench BD	VANESSA	Symbolic	50.00	x	x	x
		LLaMa3	52.28	55.56	50.0	52.63
		Ministral	59.12	69.23	45.0	54.55
	LINC	LLaMa3	24.91	40.0	20.0	26.67
		Ministral	24.91	37.5	15.0	21.43
	CoT	LLaMa3	**59.12**	57.69	75.0	65.22
		Ministral	50.00	48.15	65.0	55.32
	Direct	LLaMa3	63.69	62.5	75.0	68.18
		Ministral	22.63	20.0	40.0	26.67

Table 2: Accuracy, as well as Micro-averaged Precision, Recall, and F1 for the positive and negative classes. The small line divides grey-box and black-box models. Green: best F1 overall. Cyan: best F1 for the other category of models. Bold: best accuracy. Underlined: best accuracy for the other side.