

Neuro-Symbolic Logical Reasoning with Textual Entailment

Zacchary Sadeddine, Fabian M. Suchanek

Telecom Paris, Institut Polytechnique de Paris, France

zacchary.sadeddine@telecom-paris.fr fabian.suchanek@telecom-paris.fr

Abstract

Large Language Models can use logical deduction to answer natural language questions, but they remain black-boxes with potentially erroneous chains-of-thought. In this paper, we adapt VANESSA, a neuro-symbolic method for chain-of-thought verification, to reasoning-based question answering. VANESSA combines a logical reasoner with a neural textual entailment model to handle phrasing variations. Building on VANESSA, we develop a transparent, logic-based approach to answer natural language questions even with phrase variations. Our experiments across various datasets show our method is competitive with the state of the art, while also delivering proof trees for its answers. A demo interface allows users to interact with the system.

1 Introduction

Reasoning-based question answering (QA) 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 is also 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 even when the final answer is correct [Golovneva *et al.*, 2023; Sadeddine and Suchanek, 2025]. Purely symbolic reasoners, in contrast, can deliver a formal proof for their answer. However, they get derailed by phrase variations. In our example, they are unable to see that “is a fan of bread” has the same meaning as “likes bread”, jeopardizing the proof.

In this demo paper, we propose a neuro-symbolic method to answer such questions. Our idea is to build on VANESSA –

a method to verify the reasoning steps of LLMs [Sadeddine and Suchanek, 2025]. The central principle of VANESSA is to use textual entailment to find correspondences between phrases, e.g., to find that “is a fan of bread” entails “likes bread”. Once these correspondences have been established, VANESSA runs a symbolic prover to find the answer. The answer is accompanied by a formal proof, in which the only fallible components are the entailment steps.

In this paper, we adapt VANESSA to perform reasoning-based question answering directly. Our method produces answers that are accompanied by transparent reasoning steps and a proof tree. Our experiments on multiple logical QA benchmarks compare our adapted VANESSA to black-box and neuro-symbolic competitors, and show that VANESSA is competitive in overall performance with purely neural methods, while additionally delivering a 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 [Wei *et al.*, 2022] has further increased model performance while giving the user access to a proof, but hallucinations and formal errors still can be present [Golovneva *et al.*, 2023; Sadeddine and Suchanek, 2025]. 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 [Wang *et al.*, 2023; Fang *et al.*, 2024; Ge *et al.*, 2025; Suchanek and Luu, 2023]. 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 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, builds transparent proof trees, in which the only non-symbolic (and hence fallible) components are the textual entailment steps. Thereby, the area of distrust is reduced to a single atomic task, on which LLMs usually

ProofWriter	Accuracy	Prec	Rec	F1
★★ VANESSA symb.	88.27	97.56	83.33	89.89
VANESSA LLaMa3	65.90	63.03	78.12	69.77
VANESSA Ministral3	62.95	58.62	70.83	64.15
LINC	73.55	94.52	71.88	81.66
LINC Ministral3	68.25	94.29	68.75	79.52
CoT LLaMa3-8B	45.29	40.19	44.79	42.37
CoT Ministral3	64.72	61.68	68.75	65.02
Direct LLaMa3-8B	27.04	24.83	38.54	30.2
Direct Ministral3	58.24	57.14	62.50	59.70

ProofWriter	Accuracy	Prec	Rec	F1
★★ VANESSA symb.	39.01	96.77	23.62	37.97
VANESSA LLaMa3	84.18	85.03	98.43	91.24
★★ VANESSA Ministral3	92.72	92.59	98.43	95.42
LINC	65.87	85.09	76.38	80.5
LINC Ministral3	70.75	78.62	89.76	83.82
CoT LLaMa3-8B	68.92	74.48	85.04	79.41
CoT Ministral3	64.04	73.91	80.31	76.98
Direct LLaMa3-8B	59.16	59.38	74.8	66.2
Direct Ministral3	68.92	79.26	84.25	81.68

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	36.05	100.0	2.96	5.75
VANESSA LLaMa3	49.52	59.14	40.74	48.25
★ VANESSA Ministral3	50.48	70.59	35.56	47.29
LINC	34.12	86.84	24.44	38.14
LINC Ministral3	33.16	84.78	28.89	43.09
CoT LLaMa3-8B	55.77	57.06	74.81	64.74
CoT Ministral3	33.80	33.33	34.07	33.7
Direct LLaMa3-8B	55.92	53.3	71.85	61.2
Direct Ministral3	63.95	71.94	74.07	72.99

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	0.00	x	x	x
VANESSA LLaMa3	54.56	88.0	55.0	67.69
★ VANESSA Ministral3	61.40	83.33	62.5	71.43
LINC	29.47	84.62	27.5	41.51
LINC Ministral3	11.22	100.0	7.5	13.95
CoT LLaMa3-8B	65.97	87.1	67.5	76.06
CoT Ministral3	70.53	96.67	72.5	82.86
Direct LLaMa3-8B	31.75	38.71	30.0	33.8
Direct Ministral3	61.40	96.15	62.5	75.76

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	50.00	x	x	x
★ VANESSA LLaMa3	56.84	56.0	70.0	62.22
VANESSA Ministral3	50.00	57.14	20.0	29.63
LINC	52.28	75.0	45.0	56.25
LINC Ministral3	40.88	81.82	45.0	58.06
CoT LLaMa3-8B	52.28	51.35	95.0	66.67
CoT Ministral3	56.84	71.43	100.0	83.33
Direct LLaMa3-8B	50.00	48.65	90.0	63.16
Direct Ministral3	72.81	70.83	85.00	77.27

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	50.00	x	x	x
★★ VANESSA LLaMa3	52.28	50.0	45.0	47.37
VANESSA Ministral3	52.28	50.0	25.0	33.33
LINC	31.75	31.25	25.0	27.78
LINC Ministral3	27.19	70.0	35.0	46.67
CoT LLaMa3-8B	50.00	47.83	55.0	51.17
CoT Ministral3	45.44	33.33	15.00	20.69
Direct LLaMa3-8B	50.00	50.0	70.0	58.33
Direct Ministral3	40.88	54.17	65.00	59.09

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	50.00	x	x	x
★ VANESSA LLaMa3	52.28	55.56	50.0	52.63
VANESSA Ministral3	54.56	56.25	45.0	50.0
LINC	24.91	40.0	20.0	26.67
LINC Ministral3	31.75	66.67	30.0	41.38
CoT LLaMa3-8B	59.12	57.69	75.0	65.22
CoT Ministral3	47.72	62.07	90.00	73.47
Direct LLaMa3-8B	63.69	62.5	75.0	68.18
Direct Ministral3	43.16	65.22	75.00	69.77

ProofWriter	Accuracy	Prec	Rec	F1
VANESSA symb.	50.00	x	x	x
VANESSA LLaMa3	46.42	40.0	15.0	21.82
VANESSA Ministral3	44.04	27.27	7.5	11.76
★ LINC	54.77	61.54	40.0	48.49
LINC Ministral3	30.92	63.64	17.5	27.45
CoT LLaMa3-8B	58.35	55.56	87.5	67.96
CoT Ministral3	55.96	53.33	100.00	69.56
Direct LLaMa3-8B	50.00	x	x	x
Direct Ministral3	53.58	52.00	97.5	67.83

Background color indicates white-box, gray-box, and black-box approaches. ★ marks the best white-box/gray-box approach; ★★ marks a white-box/gray-box approach that beats even black-box approaches.

Table 1: Accuracy, as well as micro-averaged precision, recall, and F1 for the positive and negative classes.

84 perform well. Besides, an entailment step is usually easy to
85 verify manually.

86 3 VANESSA

87 VANESSA is a method to verify the reasoning within a chain-
88 of-thought [Sadeddine and Suchanek, 2025]. The input to
89 VANESSA is a context, a boolean question, and a chain-of-
90 thought that answers the question. The chain-of-thought is
91 a sequence of reasoning steps, each of which consists of one
92 or more premises and a conclusion. VANESSA then checks
93 every single reasoning step and outputs “Correct” if every step
94 is valid and every premise is grounded in the context or in
95 previous conclusions. VANESSA operates in three phases:
96 (1) a shallow symbolic parsing of the context and the ques-
97 tion, (2) an augmentation of the logical forms through textual
98 entailment, and (3) symbolic reasoning. Step (2) can be per-
99 formed symbolically by string matching (resulting in a fully
100 symbolic variant of VANESSA) or with an LLM (yielding a
101 neuro-symbolic variant, which is more robust to variations in
102 phrasing).

103 The present work adapts VANESSA to perform question
104 answering directly, without requiring a given chain-of-thought.
105 The input to our adapted method is a context, consisting of
106 simple rules and facts in natural language, and a boolean ques-
107 tion (as in the example in the introduction). The context has
108 to be self-contained, i.e., no external knowledge is needed to
109 answer the question (in particular, no mathematical knowl-
110 edge). There are three possible answers to the question: “Yes”,
111 “No” and “Unknown” (if the context does not permit a definite
112 conclusion).

113 We transform this input into a pseudo-reasoning step, which
114 has the entire context as premises, and the question (in the
115 form of an affirmative sentence) as the conclusion. Like in the
116 original VANESSA, we try to validate the reasoning step. If

this succeeds, the answer to the question is “Yes”. If it fails, 117
our adapted method then negates the conclusion and tries to 118
validate it. If that succeeds, the answer is “No”. Otherwise the 119
answer is “Unknown”. When VANESSA successfully finds 120
an answer, it automatically constructs a proof tree, which is 121
presented as an explanation supporting the answer. 122

4 Experiments 123

We evaluate our method on several logical reasoning datasets: 124
ProofWriter [Taffjord *et al.*, 2021] (“Depth 5, Open World 125
Assumption” Dev set), ProntoQA [Saparov and He, 2023] 126
(using the 100 first instances of the 4-hop Composed Random 127
set from ProntoQA-OOD [Saparov *et al.*, 2023]), FOLIO [Han 128
et al., 2022], and LogicBench [Parmar *et al.*, 2024]. The latter 129
does not consider the Open World Assumption, and thus makes 130
no distinction between “No” and “Unknown”. Hence we 131
manually relabeled negative ground truth examples as either 132
“No” or “Unknown” for the BD, CD, DD and HS subsets, and 133
subsampling those (except HS) to achieve balance between the 134
possible answers. 135

We compare several approaches: **Black-box methods** use 136
prompt-instructed LLMs – either to obtain directly an answer 137
to the question, or to obtain a chain-of-thought that answers 138
the question. **Gray-box approaches** are neuro-symbolic, i.e., 139
they deliver a formal proof, albeit with fallible components. 140
We use LINC [Olausson *et al.*, 2023] and the neuro-symbolic 141
VANESSA. Finally, **white-box approaches** are fully symbolic, 142
and we use the symbolic VANESSA. 143

Table 1 shows that, as expected, the black-box approaches 144
generally perform best. In general, the chain-of-thought 145
prompting outperforms direct prompting approach – albeit 146
only on 3 out of the 5 datasets. None of the black-box meth- 147
ods offers a proof tree. 148

Turning to the white-box approach, the symbolic VANESSA 149

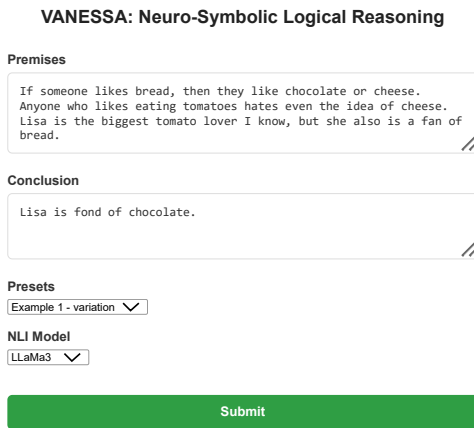


Figure 1: System Interface

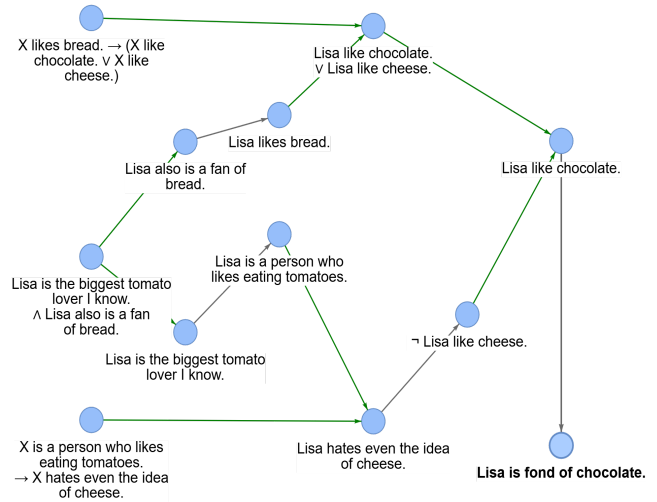


Figure 2: Proof Graph for the example

1. Input: $X \text{ likes bread} \rightarrow (X \text{ like chocolate} \vee X \text{ like cheese})$
2. Instantiation of 1: $\text{Lisa likes bread} \rightarrow (\text{Lisa like chocolate} \vee \text{Lisa like cheese})$
3. Input: $\text{Lisa is the biggest tomato lover I know} \wedge \text{Lisa also is a fan of bread}$
4. Deduction (3): $\text{Lisa also is a fan of bread}$
5. Entailment (from 4): Lisa likes bread
6. Deduction (2+5): $\text{Lisa like chocolate} \vee \text{Lisa like cheese}$
7. Input: $X \text{ is a person who likes eating tomatoes} \rightarrow X \text{ hates even the idea of cheese}$
8. Instantiation of 7: $\text{Lisa is a person who likes eating tomatoes} \rightarrow \text{Lisa hates even the idea of cheese}$
9. Deduction (3): $\text{Lisa is the biggest tomato lover I know}$
10. Entailment (from 9): $\text{Lisa is a person who likes eating tomatoes}$
11. Deduction (8+10): $\text{Lisa hates even the idea of cheese}$
12. Entailment (from 11): $\neg \text{Lisa like cheese}$
13. Deduction (6+12): $\text{Lisa like chocolate}$
14. Entailment (from 13): $\text{Lisa is fond of chocolate}$

Figure 3: Textual Proof for the example

150 also performs as expected: Whenever it delivers results, these
 151 consistently have the highest precision. It even reaches best
 152 overall accuracy on ProofWriter. However, the method falls
 153 behind on recall because of its inability to deal with phrasing
 154 variations. On the LogicBench datasets, this issue goes so
 155 far that the method fails to deliver any verdict at all, always
 156 outputting “unknown”, which results in an accuracy of 50%
 157 on subsets where half the ground truth labels are “unknown”.
 158 Among the gray-box approaches, the neuro-symbolic
 159 VANESSA consistently achieves higher accuracy than LINC.
 160 On several datasets, VANESSA beats even the black-box ap-
 161 proaches, a feat that LINC does not achieve. Overall, our ex-
 162 periments thus show that symbolic and neuro-symbolic meth-
 163 ods can compete with black-box models in terms of accuracy
 164 – while additionally providing a symbolic explanation. The
 165 neuro-symbolic VANESSA emerges as the best-performing
 166 gray-box method.

167 5 Demo

168 A demonstration of our system is available at <https://vanessa-demo.org/>, and it can be used online or downloaded
 169 for local use. The user can input premises and a conclusion for
 170 a logical reasoning problem, and run VANESSA in either sym-
 171 bolic or neuro-symbolic mode (Figure 1). The neuro-symbolic
 172 variant uses LLaMa3.2-3B. It is somewhat slow in the online
 173 interface due to computational requirements, but is usually
 174 faster when run locally.

176 When VANESSA finds a solution to the reasoning problem,
 177 the interface displays a proof tree (Figure 2). Green arrows
 178 indicate logical deduction, i.e., trustworthy steps. Gray arrows
 179 are the potentially erroneous entailment steps, which the user
 180 has to check. The interface also shows the parsed input sen-
 181 tences, the detected entailments and a linearized textual proof
 182 (Figure 3), allowing users to trace the reasoning process.

183 In the demo, users can play around with the preset examples
 184 that the GUI offers from several benchmarks. Users can also
 185 modify the examples, for example by rephrasing sentences
 186 to test the system’s robustness, adding negations or changing
 187 conclusions. Finally, they can also submit their own reasoning

problems and see if the system can give the correct response. 188

189 6 Conclusion

190 We have presented an adaptation of VANESSA for answering
 191 natural language questions. Our method capitalizes on the
 192 key principle of VANESSA, which uses a formal reasoner and
 193 bridges differences in phrasing by textual entailment. Our
 194 experiments show that our adaptation of VANESSA can com-
 195 pete with black-box models on the task of reasoning-based
 196 question answering, while also providing a proof tree. In a
 197 hands-on demo, users can play with the system, explore its
 198 capabilities, and submit their own logical riddles.

199 We hope that this work paves the way for the devel-
 200 opment of more explainable and transparent logical rea-
 201 soning systems. All code and data is available at <https://github.com/dig-team/VANESSA/tree/demo>, with a video at
 202 <https://youtu.be/K426slpTtAE>. 203

204 Acknowledgements

205 This work was partially funded by the NoRDF project (ANR-
206 20-CHIA-0012-01).

207 References

208 Nasim Borazjanizadeh and Steven T Piantadosi. Reli-
209 able reasoning beyond natural language. *arXiv preprint*
210 *arXiv:2407.11373*, 2024.

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

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

221 Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin
222 Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and
223 Asli Celikyilmaz. ROSCOE: A suite of metrics for scor-
224 ing step-by-step reasoning. In *The Eleventh International*
225 *Conference on Learning Representations*, 2023.

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

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

237 Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Ar-
238 mando Solar-Lezama, Joshua Tenenbaum, and Roger Levy.
239 LINC: A neurosymbolic approach for logical reasoning by
240 combining language models with first-order logic provers.
241 In Houda Bouamor, Juan Pino, and Kalika Bali, editors,
242 *Proceedings of the 2023 Conference on Empirical Methods*
243 *in Natural Language Processing*, pages 5153–5176, Sin-
244 gapore, December 2023. Association for Computational
245 Linguistics.

246 Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Naka-
247 mura, Man Luo, Santosh Mashetty, Arindam Mitra, and
248 Chitta Baral. LogicBench: Towards systematic evaluation
249 of logical reasoning ability of large language models. In
250 Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors,
251 *Proceedings of the 62nd Annual Meeting of the Association*
252 *for Computational Linguistics (Volume 1: Long Papers)*,
253 pages 13679–13707, Bangkok, Thailand, August 2024. As-
254 sociation for Computational Linguistics.

255 Zacchary Sadeddine and Fabian M. Suchanek. Verifying the
256 steps of deductive reasoning chains. In *Proceedings of*
257 *the 2025 Conference of the Association for Computational*

Linguistics: Human Language Technologies (Volume 2:
Findings). Association for Computational Linguistics, 2025. 258 259

Abulhair Saparov and He He. Language models are greedy
reasoners: A systematic formal analysis of chain-of-thought.
In *The Eleventh International Conference on Learning Rep-*
resentations, 2023. 260 261 262 263

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmaku-
mar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and
He He. Testing the general deductive reasoning capac-
ity of large language models using OOD examples. In
Thirty-seventh Conference on Neural Information Process-
ing Systems, 2023. 264 265 266 267 268 269

Fabian M. Suchanek and Anh Tuan Luu. Knowledge
Bases and Language Models: Complementing Forces. In
RuleML+RR invited paper, 2023. 270 271 272

Oyvind Taffjord, Bhavana Dalvi, and Peter Clark. ProofWriter:
Generating implications, proofs, and abductive statements
over natural language. In Chengqing Zong, Fei Xia, Wenjie
Li, and Roberto Navigli, editors, *Findings of the Associ-*
ation for Computational Linguistics: ACL-IJCNLP 2021,
pages 3621–3634, Online, August 2021. Association for
Computational Linguistics. 273 274 275 276 277 278 279

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. Lego-prover: Neural theorem proving with
growing libraries, 2023. 280 281 282 283 284

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma,
Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought
prompting elicits reasoning in large language models. *CoRR*,
abs/2201.11903, 2022. 285 286 287 288

Sen Yang, Xin Li, Leyang Cui, Lidong Bing, and Wai Lam.
Neuro-symbolic integration brings causal and reliable rea-
soning proofs. *arXiv preprint arXiv:2311.09802*, 2023. 289 290 291