ProofTeller: Exposing Recency Bias in LLM Reasoning and Its Side Effects on Communication (Extended Abstract)

Mayank Jobanputra[1], Alisa Kovtunova[2], Brisca Balthes[1,3], Fedor Grigoryevich Pogulskiy[2], Yifan Wang[1], Stefan Borgwardt[2] and Vera Demberg[1]

[1]Saarland University, Saarbrücken, Germany
[2]Institute of Theoretical Computer Science, TU Dresden, Germany
[3]Zuse School ELIZA, Germany

Abstract

This abstract is for a paper [1] published in 2025 in the proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, and we extend it here with a new study that closes the gap with previous research.

Large language models (LLMs) are increasingly applied in domains that demand reliable and interpretable reasoning. While formal reasoning methods can generate correct proofs, these proofs are often inaccessible to non-expert users. This raises a natural question: Can LLMs, when given a proof, faithfully interpret its reasoning and communicate it clearly? Recently, we have introduced `ProofTeller`, a benchmark that evaluates this ability across three tasks: (1) identifying key proof steps, (2) summarizing the reasoning, and (3) explaining the result in concise natural language. The benchmark covers three domains: *Biology*, *Drones*, and *Recipes*, representing scientific, safety-critical, and everyday reasoning scenarios. We find a consistent near-conclusion bias: LLMs tend to focus on steps closest to the final proof conclusion rather than on the most informative ones. A targeted human study confirms that explanations based on such steps are rated less appropriate for end users. These findings indicate that even when reasoning is provided, current LLMs face challenges in communicating key information in a useful manner, highlighting the need for LLMs that can communicate important details reliably.

Keywords

proofs, benchmark, LLM evaluation, reliability, faithfulness

1. Introduction

Large language models (LLMs) are increasingly applied in domains that demand reliable and interpretable reasoning [2, 3, 4]. However, even improved, their outputs remain unpredictable [5]. Small changes in phrasing or ordering can lead to inconsistent answers [6, 7], and explanations often appear plausible yet fail to reflect the reasoning behind the prediction [8, 9]. On the other hand, formal reasoning methods can generate correct proofs but these proofs are often inaccessible to non-expert users due to their format or excessive length [10]. Combining the two should play to each component's strengths: the proof provides correctness, and the LLM focuses on communication. Namely, if an LLM is supplied with a proof as input, the reasoning path is already complete and verifiable; the LLM's task is to interpret the proof, identify its essential steps, and communicate them clearly. This setting isolates the linguistic and communicative aspects of reasoning from the purely deductive ones, allowing us to study whether LLMs can serve as effective interfaces between formal reasoning systems and human users [11]. However, as Turpin et al. [8] observe, faithfulness in current LLMs is easily influenced by superficial cues in the input.

DL 2026: 39th International Workshop on Description Logics, July 17–19, 2026, Lisbon, Portugal
mayank@lst.uni-saarland.de (M. Jobanputra); alisa.kovtunova@tu-dresden.de (A. Kovtunova); briscab@lst.uni-saarland.de (B. Balthes); fedor_grigoryevich.pogulskiy@mailbox.tu-dresden.de (F. G. Pogulskiy); yifwang@lst.uni-saarland.de (Y. Wang); stefan.borgwardt@tu-dresden.de (S. Borgwardt); vera@lst.uni-saarland.de (V. Demberg)
https://mayankjobanputra.github.io/ (M. Jobanputra); https://lat.inf.tu-dresden.de/~alisa/ (A. Kovtunova); https://lat.inf.tu-dresden.de/~stefborg/ (S. Borgwardt)
0000-0002-8802-2401 (M. Jobanputra); 0000-0001-9936-0943 (A. Kovtunova); 0000-0003-0924-8478 (S. Borgwardt); 0000-0002-8834-0020 (V. Demberg)

These observations lead to a question we address in this work: can LLMs reliably interpret a formal proof, identify the key reasoning steps, and restate them faithfully and intelligibly for a non-expert?

To investigate this problem, we introduce `ProofTeller`[1], a benchmark that evaluates how well LLMs interpret and communicate verified proofs on three different tasks: **Task 1**: highlight the key steps leading to the proof's conclusion; **Task 2**: summarize the overall reasoning faithfully; and **Task 3**: explain the result in a concise target message to a specific user group.

In addition, in Section 3 we extend [1] with an additional experiment that compares the quality of LLM outputs when the input consists of a full proof versus when it only consists of a justification [12, 13, 14], i.e., a minimal subset of ontology axioms entailing a given consequence. Proofs, justifications, LLM outputs, and human ratings are available online [15].

2. The ProofTeller Benchmark

Our proofs [16, 17, 18, 19, 20, 21] come from three distinct domains:

- *Biology*. Our first domain is based on the Cell Line Ontology,[2] a community-driven ontology developed to standardize and integrate cell line information and support computer-assisted reasoning. *Scenario*: a 10th-grade student learning about the characteristics of various cells.
- *Recipe*. We created a food and recipe ontology [22], a formal semantic model to represent and reason about culinary recipes, dietary restrictions, and allergen content. This ontology builds upon established recipe and food ontologies [23, 24]. *Scenario*: a 6th-grade student learning about food allergens and and dietary classifications (vegan, vegetarian, non-vegetarian).
- *Drone*. The drone ontology is a DatalogMTL program [25] also presented in Borgwardt et al. [26] that models complex situations occurring during a drone flight. *Scenario*: a drone pilot monitoring an autonomous drone, needing help identifying (potential) critical situations.

For the *Biology* and *Recipe* domains, we use the description logics (DLs) $\mathcal{EL}$ and $\mathcal{ALC}$ [27]. For the *Drone* domain, we use the temporal logic DatalogMTL [28]. Proofs are automatically generated using existing reasoners: Elk [29, 30] for $\mathcal{EL}$, Lethe [31, 18] for $\mathcal{ALC}$, and MeTeoR [32] for DatalogMTL, with Evee [33] used to extract size-minimal proofs. For each target domain, we randomly sampled 50 proofs for annotation, except for *Recipe* for which we intentionally keep 10 $\mathcal{ALC}$ proofs. As a result, we obtained proofs from the *Biology* domain ranging from 4 to 29 inferences (mean = 17.5) of at most 2 premises per inference, from *Recipe* ranging from 11 to 63 inferences (mean = 20.5), of at most 3 premises per inference, and from *Drone* ranging from 9 to 75 inferences (mean = 24.6) of at most 6 premises per inference. To form a baseline, we recruited two graduate-level students to complete **Task 1-3** given a verified proof as input.

Key Results In the following we report the performance of nine LLMs (GPT 4.1 mini, GPT 4o mini [34], SmolLM3 3B [35], Llama 3.1 8B, Llama 3.3 70B [36], Mistral 3.2 24B, Magistral 24B [37], Qwen3 8B, and Qwen3 32B [38]) on **Tasks 1-3**.

For **Task 1**, to quantitatively evaluate the choice of the key steps, we measure the average depth of the selected proof steps within the reference proof. We represent each proof as a Directed Acyclic Graph (DAG), $G = (V, E)$, where vertices $v \in V$ are the inferences, a directed edge $(u, v) \in E$ indicates that the conclusion of v is a premise for u, and $\texttt{depth}(v)$ is the length of the longest path from the proof's final conclusion v_{root} to v. For a given set of selected steps $S = \{s_1, \ldots, s_k\}$, we calculate the mean normalized depth $\bar{d}(S)$:

$$\bar{d}(S) = \frac{1}{k} \sum_{i=1}^{k} \frac{\texttt{depth}(s_i)}{\max_{u \in V} \texttt{depth}(u)}.$$

[1] https://github.com/mayankjobanputra/ProofTeller
[2] https://obofoundry.org/ontology/clo.html

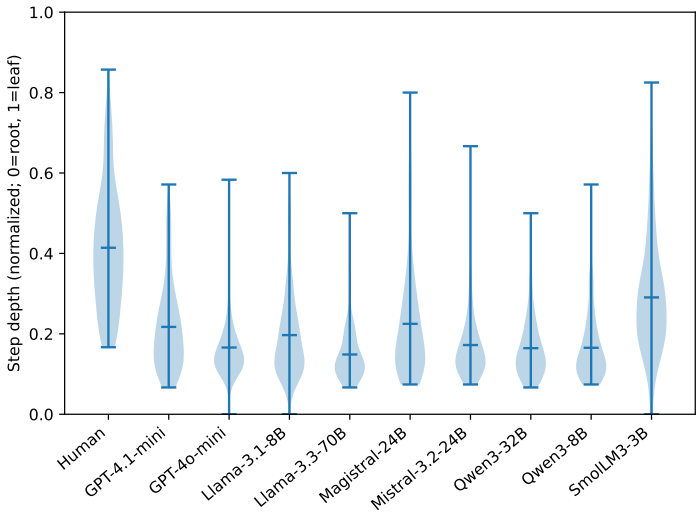

Figure 1: Most contributing steps distribution (averaged across domains)

This ensures that $\bar{d}(v) \in [0, 1]$, where 0 corresponds to only the root (final conclusion) being selected, and 1 to selecting among the deepest leaf nodes (assertions and axioms/rules).

Figure 1 presents the distributions of the mean normalized step depth for most contributing steps identified by humans and LLMs across all domains. Human-identified steps exhibit the highest mean step depth (≈ 0.4) and the widest distribution, covering nearly the entire normalized range. This indicates that humans do not adhere to a single fixed strategy. Instead, they flexibly select steps from all levels of the proof graph, from high-level conclusions to foundational premises. In contrast, most LLMs display a strong and consistent bias toward low-depth steps, indicating a preference for inferences that are very close to the final conclusion.

For **Tasks 2-3**, we evaluated similarity between the LLM-generated text and the human annotations using standard metrics such as BLEU [39], ROUGE-L [40], chrF++ [41], and BERTScore [42] (see figures in the original work [1] in the appendix). The highest scores are achieved by Llama-3.1-8B, GPT-4o-mini, and Mistral-3.2-24B. In contrast, LLMs like GPT-4.1-mini and SmolLM3-3B generally rank lower on these generation tasks compared to their performance on the step selection task.

To provide a more robust assessment, we also conducted a human evaluation study. To optimize the effort, we selected a subset of models for comparison against the human baseline: the two top performers according to automated metrics (Mistral-3.2-24B and Llama-3.1-8B), a larger model from the same family (Llama-3.3-70B), and a recent proprietary model (GPT-4.1-mini). To carry out the evaluation, we recruited five human annotators to rate the outputs for both tasks.

The human evaluation statistics are presented in Table 1. The table indicates qualitative differences between the LLMs and Human annotators. Based on this evaluation, GPT-4.1-mini and Mistral-3.2-24B are the best-performing LLMs overall and they score higher compared to the Llama family LLMs. For the

Table 1
Averaged human evaluation scores for the LLMs output quality across all domains and annotators

	Summary criteria (**Task 2**)				Target Message criteria (**Task 3**)		
Model Name	Conciseness	Coverage	Faithfulness	Readability	Appropriateness	Coverage	Faithfulness
Llama-3.1-8B	4.07	3.65	3.76	4.05	3.54	3.69	3.67
Mistral-3.2-24B	4.01	4.63	3.86	3.83	3.85	4.19	4.11
GPT-4.1-mini	3.79	4.60	4.00	3.57	3.89	4.33	4.24
Llama-3.3-70B	4.14	3.84	4.00	3.84	3.62	3.46	4.44
Human	**4.55**	**4.80**	**4.80**	**4.35**	**4.59**	**4.47**	**4.96**

Table 2
Human evaluation scores comparing justification-based and proof-based outputs across all LLMs and annotators

Domain	Tasks	Proof better	Justification better	Tied / no majority
Biology	40	17.5%	45.0%	37.5%
Recipe	40	20.0%	17.5%	62.5%
Drone	40	82.5%	7.5%	10.0%
Overall	120	40.0%	23.3%	36.7%

summary task (**Task 2**), GPT-4.1-mini and Mistral-3.2-24B obtained the highest scores for the Coverage criteria, but they fall short on the Conciseness criteria. This indicate the verbose nature of these LLMs. In the target message task (**Task 3**), GPT-4.1-mini achieves the highest scores.

3. Comparison to Justifications

In addition to the results from [1], we investigated whether including the information from the full proof actually improves the generation of concise explanations, compared to approaches that only provide justifications as input [43]. To do this, we re-ran **Task 3** with the best performing LLMs in Table 1, GPT-4.1-mini and Mistral-3.2-24B, with slightly adapted prompts for justification proofs, because they include no intermediate proof steps between the assertions and the conclusion. We then asked seven human experts to compare, for each LLM, whether the proof-based or justification-based explanation was better (see the annotator interface in Figure 2 of the appendix), following these instructions:

A good target message must (i) explain what happened, i.e., state the key conclusion clearly (e.g. "Warning! Drone battery is critically low."); (ii) explain why, i.e., provide a reason (e.g. "...because battery level drops to 2% before the goal is reached"); (iii) be concise, i.e., ideally ≤ 20 *words, cover only what matters.*

Using a five-point scale ("A is much better than B", "A is better than B", "A and B are equally good/bad", "B is better than A", and "B is much better than A"), the annotators produced the results in Table 2.

We can see that using proofs confers no advantage over justifications in the simple *Recipe* domain, and actually results in worse explanations in the *Biology* domain. One possible explanation is that high-level justifications-based outputs may be better aligned with how laypeople see complex biological reasoning. For example, the justification-based message *"D-392MG cells are cancer cell lines because they are immortal and model glioblastoma, a type of cancer."* was rated as "much better" by four annotators, and one said it is "better" than the proof-based explanation *"D-392MG cells are cancer models and immortal, making them cancer cell lines."* This suggests that explicitly naming a specific cancer type (glioblastoma) rather than using the generic phrase "cancer models" adds perceived credibility to the sentence, even though the two statements are semantically similar. It may also related to the observed bias of LLMs to focus on later steps in the proof, where the specific type of cancer has already been replaced by more generic concepts.

In contrast, in the more complex *Drone* domain, the additional information in the proofs was often necessary to formulate a clearer explanation, by shifting the focus away from details in the input data. For example, the proof-based message *"Warning! Drone d is at risk of crash due to proximity to a static rock ahead."* was rated "much better" by four annotators, and one said it is "better" than the justification-based message *"Warning! Drone near rock at position x9y13 with high speed at current time"*.

In both examples, the proofs resulted in more generic messages. Depending on the domain and scenario, this could be beneficial, for instance, when a complex event is being summarized concisely as in *Drone*, or it could produce a bland, less informative message as in *Biology*.

4. Conclusion

Paper [1] introduced `ProofTeller`, a benchmark to evaluate LLM reliability in explaining formal proofs via key step identification, summarization, and targeted message generation. Our experiments reveal reliability gaps: LLMs consistently favor low-depth steps near the conclusion, whereas humans select steps from all over the proof, suggesting a lack of holistic understanding. Human evaluation further highlights qualitative deficiencies in the output. Additionally, the relative performance of proof-based versus justification-based inputs was not uniform: it depended significantly on the domain.

Acknowledgments

The authors thank the survey participants, Christian Alrabbaa, Pratistha Kansakar, Oliver Fernandez Gil, and Philipp Max Hermann, for their contributions.

In the original paper [1] the authors extended gratitude to Saumil Shah, Leixin Zhang, Zaynab Reza, Pratistha Kansakar and Zhenyu Feng for their help in the human evaluation study. We would like to thank Yash Sarrof, Dongqi Liu, Soyoung Oh, and Blerta Veseli for discussions and feedback on the draft. This work is funded by Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science. Brisca Balthes was further supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. Yifan Wang was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914. The authors gratefully acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner).

Declaration on Generative AI

During the preparation of this work, the authors used DeepSeek-V3 in order to: Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] M. Jobanputra, A. Kovtunova, B. Balthes, F. G. Pogulskiy, Y. Wang, S. Borgwardt, V. Demberg, Proofteller: Exposing recency bias in LLM reasoning and its side effects on communication, in: K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, D. P. Singh (Eds.), Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025, The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025, pp. 1439–1462. URL: https://aclanthology.org/2025.ijcnlp-long.80/.

[2] Y. Bang, Z. Ji, A. Schelten, A. Hartshorn, T. Fowler, C. Zhang, N. Cancedda, P. Fung, Hallulens: Llm hallucination benchmark, arXiv preprint arXiv:2504.17550 (2025).

[3] C. Cui, Y. Ma, Z. Yang, Y. Zhou, P. Liu, J. Lu, L. Li, Y. Chen, J. H. Panchal, A. Abdelraouf, et al., Large language models for autonomous driving (llm4ad): Concept, benchmark, experiments, and challenges, arXiv preprint arXiv:2410.15281 (2024).

[4] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, M. Portmann, Flowtransformer: A transformer framework for flow-based network intrusion detection systems, Expert Systems with Applications 241 (2024) 122564.

[5] Y. Abbasi Yadkori, I. Kuzborskij, A. György, C. Szepesvari, To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty, Advances in Neural Information Processing Systems 37 (2024) 58077–58117.

[6] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=RIu5lyNXjT.

[7] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, Transactions of the Association for Computational Linguistics 9 (2021) 1012–1031. URL: https://doi.org/10.1162/tacl_a_00410. doi:10.1162/tacl_a_00410.

[8] M. Turpin, J. Michael, E. Perez, S. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, Advances in Neural Information Processing Systems 36 (2023) 74952–74965.

[9] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, arXiv preprint arXiv:2506.06941 (2025).

[10] E. F. de Lima, C. Lingenfelder, Presentation of proofs in modal natural deduction, J. Log. Comput. 10 (2000) 527–572. URL: https://doi.org/10.1093/logcom/10.4.527. doi:10.1093/LOGCOM/10.4.527.

[11] N. Kassner, O. Tafjord, H. Schütze, P. Clark, BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8849–8861. URL: https://aclanthology.org/2021.emnlp-main.697/. doi:10.18653/v1/2021.emnlp-main.697.

[12] F. Baader, B. Suntisrivaraporn, Debugging SNOMED CT using axiom pinpointing in the description logic $\mathcal{EL}^+$, in: Proc. of the 3rd Conference on Knowledge Representation in Medicine (KR-MED'08): Representing and Sharing Knowledge Using SNOMED, volume 410 of *CEUR-WS*, 2008.

[13] S. Schlobach, R. Cornet, Non-standard reasoning services for the debugging of description logic terminologies., in: G. Gottlob, T. Walsh (Eds.), Proc. of the 18th Int. Joint Conf. on Artificial Intelligence (IJCAI 2003), Morgan Kaufmann, Acapulco, Mexico, 2003, pp. 355–362.

[14] M. Horridge, Justification Based Explanation in Ontologies, Ph.D. thesis, University of Manchester, UK, 2011. URL: https://www.research.manchester.ac.uk/portal/files/54511395/FULL_TEXT.PDF.

[15] M. Jobanputra, A. Kovtunova, B. Balthes, F. G. Pogulskiy, Y. Wang, S. Borgwardt, V. Demberg, Supplementary material for the experiment reported in "ProofTeller: Exposing recency bias in LLM reasoning and its side effects on communication (extended abstract)", 2026. URL: https://doi.org/10.5281/zenodo.19765891. doi:10.5281/zenodo.19765891.

[16] M. Horridge, B. Parsia, U. Sattler, Justification oriented proofs in OWL, in: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I, 2010, pp. 354–369. doi:10.1007/978-3-642-17746-0_23.

[17] Y. Kazakov, P. Klinov, A. Stupnikov, Towards reusable explanation services in protege, in: A. Artale, B. Glimm, R. Kontchakov (Eds.), Proc. of the 30th Int. Workshop on Description Logics (DL'17), volume 1879 of *CEUR Workshop Proceedings*, 2017. URL: http://www.ceur-ws.org/Vol-1879/paper31.pdf.

[18] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Finding small proofs for description logic entailments: Theory and practice, in: E. Albert, L. Kovacs (Eds.), LPAR23. LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning, volume 73 of *EPiC Series in Computing*, EasyChair, 2020, pp. 32–67. doi:10.29007/nhpp.

[19] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, On the complexity of finding good proofs for description logic entailments, 2020. URL: http://ceur-ws.org/Vol-2663/paper-1.pdf.

[20] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Finding good proofs

for description logic entailments using recursive quality measures, 2021. doi:10.1007/978-3-030-79876-5_17.

[21] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Combining proofs for description logic and concrete domain reasoning, in: A. Fensel, A. Ozaki, D. Roman, A. Soylu (Eds.), Rules and Reasoning - 7th International Joint Conference, RuleML+RR 2023, Proceedings, volume 14244 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 54–69. doi:10.1007/978-3-031-45072-3_4.

[22] S. Borgwardt, A. Kovtunova, Food and recipe ontology: Semantic model for recipes, allergen labelling, and dietary classification, 2025. URL: https://doi.org/10.5281/zenodo.16411319. doi:10.5281/zenodo.16411319.

[23] M. Qi, Y. Neeman, F. Eckhardt, K. Blissett, WhatToMake: a semantic web application for recipe recommendation, Rensselaer Polytechnic Institute (2018). URL: https://tetherless-world.github.io/ontology-engineering/oe2020/example/files/samplepublication.pdf.

[24] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, W. W. L. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, npj Science of Food 2 (2018). doi:10.1038/s41538-018-0032-6.

[25] S. Borgwardt, A. Kovtunova, Drone program for critical situation detection over sensor data streams, 2025. URL: https://doi.org/10.5281/zenodo.15411134. doi:10.5281/zenodo.15411134.

[26] S. Borgwardt, V. Demberg, M. Jobanputra, A. Kovtunova, D. Nhu, Explaining critical situations over sensor data streams using proofs and natural language, in: L. Giordano, J. C. Jung, A. Ozaki (Eds.), Proceedings of the 37th International Workshop on Description Logics (DL 2024), Bergen, Norway, June 18-21, 2024, volume 3739 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3739/paper-2.pdf.

[27] F. Baader, I. Horrocks, C. Lutz, U. Sattler, An Introduction to Description Logic, Cambridge University Press, 2017. doi:10.1017/9781139025355.

[28] S. Brandt, E. G. Kalayci, V. Ryzhikov, G. Xiao, M. Zakharyaschev, Querying log data with metric temporal logic, J. Artif. Intell. Res. 62 (2018) 829–877. doi:10.1613/jair.1.11229.

[29] Y. Kazakov, M. Krötzsch, F. Simancik, The incredible ELK – from polynomial procedures to efficient reasoning with $\mathcal{EL}$ ontologies, J. Autom. Reasoning 53 (2014) 1–61. doi:10.1007/s10817-013-9296-3.

[30] Y. Kazakov, P. Klinov, Goal-directed tracing of inferences in EL ontologies, in: P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. Groth, N. F. Noy, K. Janowicz, C. A. Goble (Eds.), The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II, volume 8797 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 196–211. doi:10.1007/978-3-319-11915-1_13.

[31] P. Koopmann, LETHE: Forgetting and uniform interpolation for expressive description logics, KI - Künstliche Intelligenz (2020). doi:10.1007/s13218-020-00655-w.

[32] D. Wang, P. Hu, P. A. Walega, B. C. Grau, MeTeoR: Practical reasoning in datalog with metric temporal operators, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, AAAI Press, 2022, pp. 5906–5913. doi:10.1609/AAAI.V36I5.20535.

[33] C. Alrabbaa, S. Borgwardt, T. Friese, P. Koopmann, J. Méndez, A. Popovic, On the eve of true explainability for OWL ontologies: Description logic proofs with evee and evonne, in: O. Arieli, M. Homola, J. C. Jung, M. Mugnier (Eds.), Proceedings of the 35th International Workshop on Description Logics (DL 2022) co-located with Federated Logic Conference (FLoC 2022), Haifa, Israel, August 7th to 10th, 2022, volume 3263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3263/paper-2.pdf.

[34] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[35] E. Bakouch, L. Ben Allal, A. Lozhkov, N. Tazi, L. Tunstall, C. M. Patiño, E. Beeching, A. Roucher, A. J. Reedi, Q. Gallouédec, K. Rasul, N. Habib, C. Fourrier, H. Kydlicek, G. Penedo, H. Larcher, M. Morlon, V. Srivastav, J. Lochner, X.-S. Nguyen, C. Raffel, L. von Werra, T. Wolf, SmolLM3: smol,

multilingual, long-context reasoner, https://huggingface.co/blog/smollm3, 2025.
[36] A. Grattafiori, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). URL: https://arxiv.org/abs/2407.21783.
[37] A. Rastogi, et al., Magistral, arXiv preprint arXiv:2506.10910 (2025). URL: https://arxiv.org/abs/2506.10910.
[38] A. Yang, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388.
[39] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.
[40] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[41] M. Popović, chrF++: words helping character n-grams, in: O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, J. Kreutzer (Eds.), Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 612–618. URL: https://aclanthology.org/W17-4770/. doi:10.18653/v1/W17-4770.
[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020.
[43] E. Chang, A. Kovtunova, S. Borgwardt, V. Demberg, K. Chapman, H. Yeh, Logic-guided message generation from raw real-time sensor data, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, European Language Resources Association, 2022, pp. 6899–6908. URL: https://aclanthology.org/2022.lrec-1.745.

A. Annotator Interface

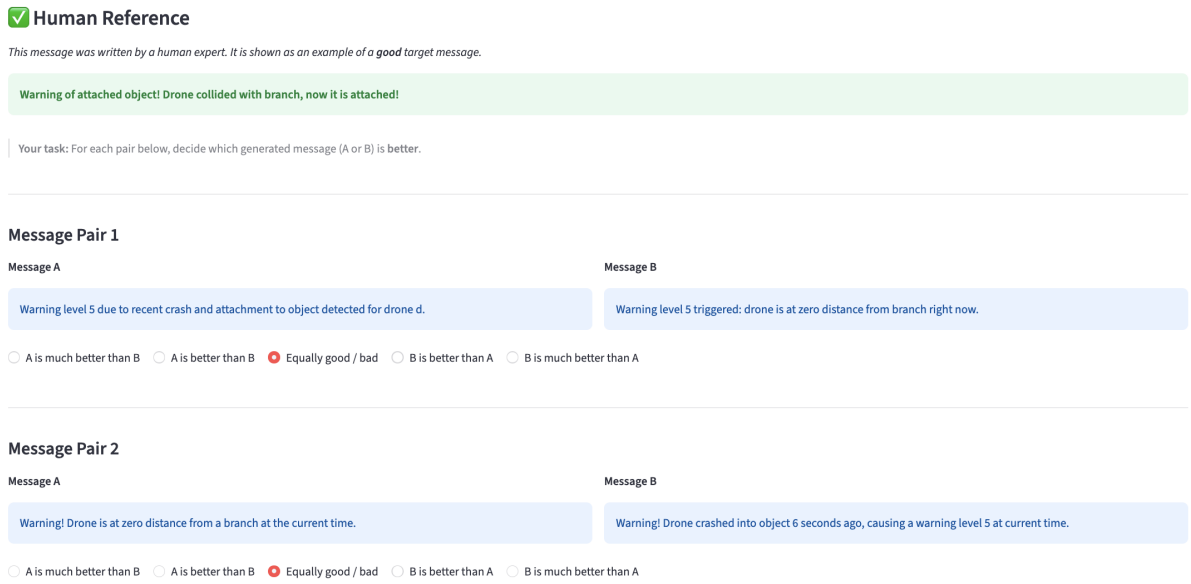

Figure 2: Annotator Interface for Comparing Target Messages Generated with Proofs and with Justifications