A Comparative Study on the Utility of Natural Language explanations for Enhancing Language Models Reasoning Performance

Anonymous ACL submission

Abstract

Natural language explanations (NLEs) are widely used to communicate model reasoning to humans, but they may also serve as effective control signals for improving model performance. In this paper, we present the first comprehensive evaluation of NLEs as prompts in in-context learning (ICL), comparing human-annotated, self-generated, and LLM-generated NLEs across five reasoning benchmarks and three instruction-tuned models (Llama 3 8B, Llama 3 70B, GPT-4o-mini). Our preliminary results show that LLM-generated explanations, especially those from GPT-4o-mini, yield the highest gains across tasks. We further plan to measure how the faithfulness of self-explanations strongly correlates to its utility, and if models retain partial robustness even when rationales are randomly mismatched or adversarially swapped.

1 Introduction

Natural language explanations (NLEs), also called rationales, have become a key mechanism for enhancing the interpretability and transparency of language models (LMs). As free-text justifications for model predictions, NLEs are widely used to communicate model reasoning to users. Beyond interpretability, however, NLEs may also serve a functional role: recent work suggests that providing explanations during inference can improve model performance.

Human-annotated rationales are often considered a gold standard, but they are expensive, slow to obtain, and subject to annotation bias and inconsistency (Yao et al., 2023; Hartmann et al., 2022). An alternative is to generate explanations automatically, either via self-explanations, in which the model justifies its own predictions, or by prompting an auxiliary LLM to generate rationales (Mishra et al., 2024; Wang et al., 2025; Wei Jie et al., 2024).

In parallel, ICL has emerged as one of the key LLM capabilities, enabling task adaptation via example-driven prompts without parameter updates (Liu et al., 2023). While ICL has shown strong performance on reasoning tasks, the impact of explanations within few-shot prompts remains underexplored. It is unclear whether different types of NLEs—human-annotated, self-generated, or LLM-generated—differ in their ability to improve model predictions when used as in-context exemplars. Furthermore, little is known about the importance of explanation quality, or how models behave when exposed to irrelevant or misleading rationales.

Objective. This work investigates the *predictive utility* of NLEs in ICL, that is, their ability to improve downstream performance when included in few-shot prompts. We systematically compare three NLE types (human-annotated, self-generated, and LLM-generated) across five reasoning benchmarks and three models (Llama 3 8B, Llama 3 70B, GPT-4o-mini). We also examine how explanation quality, quantified via faithfulness, affects performance, and how models respond to mismatched or adversarial rationales.

2 Background and Related Work

Explainable datasets and Human-annotated rationales Explainable NLP has grown, producing datasets with human-annotated explanations across various tasks (Zhao et al., 2024; Luo et al., 2024). These rationales guide training and evaluation but are costly, slow, and sometimes inconsistent (Hartmann et al., 2022). Additionally, human explanations do not always improve model performance and may only benefit specific models (Yao et al., 2023).

LLM-generated NLEs Due to the limitations of human-annotated explanations, recent research has explored using LLMs to generate NLEs (Mishra et al., 2024). Compared to traditional post-hoc feature attribution methods, NLEs provide human-readable justifications, which can enhance transparency and user understanding. Self-explanations, where models justify their own outputs, are also studied, though their faithfulness is debated. Yet, it remains unclear if LLM-generated explanations enhance downstream tasks in ICL (Madsen et al., 2024).

Leveraging Explanation to Improve Performance of LMs In a related line of work, recent studies have explored the use of explanations in ICL to enhance the reasoning capabilities of LLMs where earlier works used used costly post-hoc methods (Krishna et al., 2023). Recent methods automate rationale generation but focus on small models and don't fully assess explanation quality (Bhan et al., 2024). Our study compares human-, self-, and LLM-generated NLEs in both small and large models, exploring how explanation quality and selection affect reasoning in ICL.

3 Comparative Study Setup

Figure 1 in Appendix A.1 summarizes our framework which consists of four main steps.

Few-shot Samples Selection We follow prior work that emphasizes choosing misclassified examples to help the language model avoid similar errors on the test set (Krishna et al., 2023; Bhan et al., 2024). We adopt the n-shot sampling strategy of Bhan et al. (2024), in which error samples for a given model f are those misclassified by f in a zero-shot setting.

NLE Generation We compare three explanation types: self-NLE by the evaluated model post-prediction, human-annotated NLEs from explainable datasets, and LLM-generated NLEs created independently by other language models. This allows us to test their impact on model reasoning in ICL setting.

NLE Selection We explore selecting explanations randomly, by highest faithfulness, or by lowest faithfulness to study how explanation quality affects performance. We also test robustness by replacing rationales with random or out-of-distribution explanations from other datasets, focusing on LLM-generated rationales. This comprehensive setup helps assess the influence of explanation types and selection on model reasoning.

4 Experimental Setup

We evaluate reasoning performance on five datasets: two with human-annotated rationales—ECQA (Aggarwal et al., 2021), an extension of CommonsenseQA (Talmor et al., 2019) requiring commonsense justification , and e-SNLI (Camburu et al., 2018), a premise–hypothesis entailment dataset—and three Big-Bench-Hard (Suzgun et al., 2023) tasks including sarcasm detection (Snarks), causal reasoning (Causal Judgment), and Boolean logic evaluation. Our experiments use instruction-tuned autoregressive LLMs of varying sizes: GPT-4o-mini (Hurst et al., 2024), Llama-8B, and Llama-70B (Grattafiori et al., 2024), accessed via OpenAI and Together AI APIs, all employing a 6-shot few-shot prompting setup. For generating NLEs, we utilize GPT-4o-mini and o3-mini as explainer models, while self-explanations are generated by the evaluation models themselves. We assess the faithfulness of self-NLEs with the LExT metric (Shailya et al., 2025) Baseline comparisons include Zero-Shot prompting, Few-Shot, and Auto-CoT's (Zhang et al., 2023). To ensure reproducibility, we fix the temperature at 0 and random seeds across all experiments, and report averages over five runs for most setups.

5 Preliminary Results

Preliminary results show that LLM-NLEs, especially from GPT-4o-mini, consistently improve reasoning accuracy across models and datasets, often outperforming human and self-NLEs. High-faithfulness explanations boost performance, while low-faithfulness ones harm it, highlighting faithfulness as key. Models remain robust even with random or mismatched explanations, relying mainly on task instructions.

6 Conclusion and Outlook

Our findings suggest that NLEs can be leveraged as practical mechanisms for steering model behavior. Our results show that LLM-NLEs from 4o-mini and o3-mini can outperform self-NLEs and human-NLEs in improving the LLM reasoning performance in the ICL setting. This opens up a scalable, model-agnostic pathway for enhancing LLM performance. We plan to finalize our experiments and report final results, and to support our quantitative results with a qualitative analysis.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2024. Self-AMPLIFY: Improving small language models with self post hoc explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10974–10991, Miami, Florida, USA. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Preprint*, arXiv:1812.01193.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jan Hartmann and 1 others. 2022. Survey on how human explanations improve model learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1000–1010.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. In *Advances in neural information processing systems*, volume 36, page 65468–65483. Curran Associates, Inc. Citation Key: NEURIPS2023$_c e65173b$.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9).

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Aditi Mishra, Sajjadur Rahman, Kushan Mitra, Hannah Kim, and Estevam Hruschka. 2024. Characterizing large language models as rationalizers of knowledge-intensive tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8117–8139, Bangkok, Thailand. Association for Computational Linguistics.

Krithi Shailya, Shreya Rajpal, Gokul S Krishnan, and Balaraman Ravindran. 2025. Lext: Towards evaluating trustworthiness of natural language explanations. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1565–1587, New York, NY, USA. Association for Computing Machinery.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2025. Cross-refine: Improving natural language explanation generation by learning in tandem. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1150–1167, Abu Dhabi, UAE. Association for Computational Linguistics.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023. Are human explanations always helpful? towards objective evaluation of human natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14698–14713, Toronto, Canada. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and

Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

A Appendix

A.1 Overview of Experimental Setup

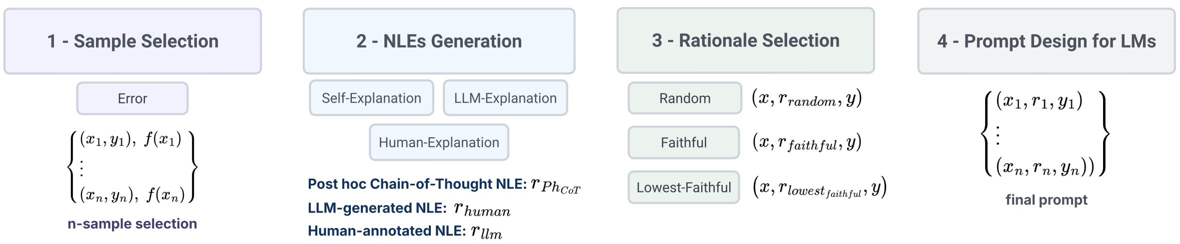

Figure 1: Overview of Experimental Setups. Self-explanation and LLM-explanation follow a four-step process aimed at gen- erating rationales to improve a language model (LM) reasoning performance in an ICL setting: (1) Samples are selected based on error selection strategy and one of three explanation selection methods: random, most faithful, or lowest faithful. (2) In the next step, NLEs are generated through: the self-explanation setup, where rationales are generated by the evaluation model itself using a post hoc explanation method (Ph-CoT) after prediction, or the LLM-explanation setup, where rationales are generated by 4o-mini and o3-mini, or human-annotated rationales are selected. (3) Rationales, and thus their respective (x, y) pairs are either randomly selected, or based on the highest faithful rationales or the lowest-faithful rationales. (4) Finally, the final ICL prompt is constructed using the selected samples and their associated rationales.