Explainable Artificial Intelligence and Reasoning in the Context of Large Neural Network Models

Stefanie Krause[1]

[1]Harz University of Applied Sciences, Friedrichstraße 57-59, Wernigerode, 38855, Germany

Abstract

The current generation of artificial intelligence (AI) systems offers tremendous benefits, but their effectiveness is limited by the inability of the machine to explain its decisions and actions to users. My dissertation will delve into the subject of explainable AI (XAI), exploring its various aspects and implications. I will focus on post-hoc local explainability of large AI models to provide human-readable explanations making users understand the automated decision-making of complex models. I aim to evaluate current large language models (LLMs) like ChatGPT or Llama on explainability and its implications, e.g., on education. My thesis also focuses on questions in the field of reasoning with LLMs, since reasoning is fundamental in LLMs for enhancing their understanding and generation of text, improving problem-solving capabilities, and facilitating natural human interaction. However, reasoning is a very difficult task for a computer and the capacities of LLMs regarding different reasoning tasks are not yet fully examined.

Keywords

explainable AI, reasoning, large language models, education

1. Motivation

Over the past few years, we have witnessed significant achievements through the utilization of large neural network models and deep learning methodologies. Nevertheless, these models operate as black boxes, requiring explanations to cause human trust in AI responses. In Figure 1 the comparison between the traditional AI decision process with a black-box model and a new process with explainability is visualized. The user of AI systems has to understand what is going on in the black-box model with an explanation interface. The aim is to make AI decisions explainable, while the challenge is that abstract algorithms find patterns in large, complex and high-dimensional data and we currently have little understanding of how this happens. At the same time, there are often problems with discrimination and bias or the Clever Hans Phenomenon, where models do not learn what cause and effect are, but only correlations [2]. These problems can be detected with the help of explainable AI. XAI can be defined as an AI that produces details or reasons to make its functioning clear or easy to understand [3]. I will focus on the explainability of AI models with the goal to provide human-readable, local explanations to make users understand the automated decision-making of complex models.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

skrause@hs-harz.de (S. Krause)

https://www.hs-harz.de/forschung/promotion/aktuelle-promotionsvorhaben/stefanie-krause/ (S. Krause)

0000-0002-1271-7514 (S. Krause)

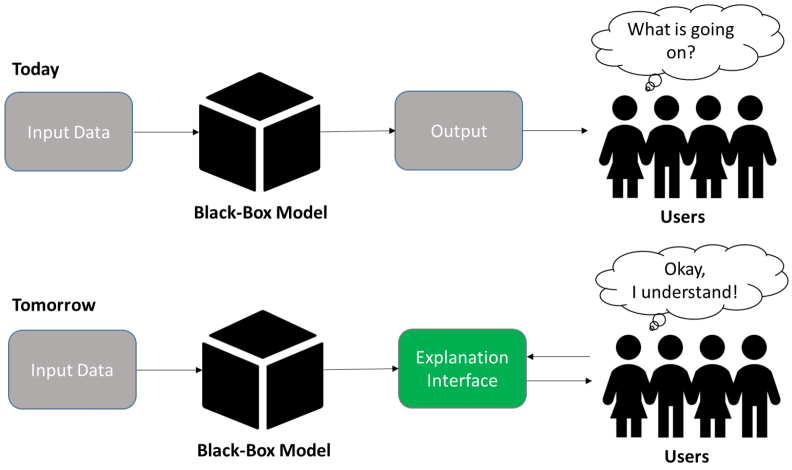

Figure 1: Comparison of a traditional AI decision process on top versus a new process with explanations given to the user. Own illustration created on basis of [1, Figure 3].

Reasoning stands as a core component of human intelligence, essential for problem-solving, making decisions, and critical thinking. Recently, advancements in large language models have made significant progress in the field of NLP, suggesting that these models might possess reasoning capabilities, especially as they increase in size [4]. Nonetheless, the full capacity of LLMs to reason effectively remains a subject of ongoing debate, which I want to explore further.

2. Related Work

At the moment LLMs like ChatGPT are dominating AI and achieved remarkable results in various tasks. In the education sector for example, students use this new technology for assignment writing among other tasks [5]. The main advantage of LLM is that one can easily generate natural language explanations for any QA task. We aim for explainability without a loss in performance and maybe even improve performance by using explanations during the training [6]. Explanations are a way to verbalize the reasoning that the models learn during training [6]. Rajani at al. developed a new dataset called Common Sense Explanations (CoS-E) and the Commonsense Auto-Generated Explanations (CAGE) framework and improved the accuracy of language models for commonsense reasoning [6]. Commonsense reasoning is a difficult challenge for a computer to handle [7]. There is a big gap between the logical approach with deductive reasoning and the inductive, associative, and empirical nature of human reasoning, rooted in past experiences. Intriguingly, LLMs lack explicit semantic knowledge, grammatical structures, or logical rules essential for explicit reasoning, not to mention large-scale ontologies found in logical knowledge bases like Adimen-SUMO [8]. A potential solution could lie in training neural networks to explicitly learn reasoning, possibly by focusing on certain sentence forms as in syllogistic reasoning may be implemented with neural-symbolic cognitive reasoning by specifically structured neural networks [9, 10]. LLMs can further be utilized to translate a natural language problem into a symbolic formulation [11]. A novel prompting technique,

known as chain-of-thought prompting, has emerged recently. This method aids LLMs in tackling reasoning challenges by directing them to generate a series of intermediate steps before providing the final answer [12]. [13] recently enhanced previous work by presenting REFINER, a system designed to enhance LLMs by training them to produce intermediate reasoning steps through interaction with a critic model that offers automated feedback on the reasoning process.

3. Research Questions

According to the previous context and motivation the primary question of my research is: **To what extent can users and developers benefit from post-hoc local explainability of large AI models?**

In a recent paper, I answered the two following research questions:

- Can LLMs like ChatGPT handle commonsense reasoning in question answering tasks with near-human-level performance?
- Are LLMs like ChatGPT able to generate good, human-understandable explanations for their decisions?

In further research, I aim to study the syllogistic reasoning abilities of LLMs as deductive reasoning is very different from the inductive reasoning that is used for commonsense reasoning. Moreover, the goal is to analyse the impact of few-shot learning and chain-of-thought as well as the potential of Retrieval-Augmented Generation (RAG). RAG offers a way to optimize the output of an LLM with specific information without changing the underlying model itself.

4. Research Approach

My objective is to rigorously test various LLMs across a carefully chosen set of reasoning tasks, aiming to compare these models' performances against that of humans. Furthermore, I plan to examine explanations of LLMs for reasoning tasks firstly by humans and secondly with an automated evaluation mechanism employing diverse scoring metrics to assess the interpretability and coherence of these models' reasoning capabilities. This comprehensive evaluation strategy seeks to illuminate the strengths and inadequacies of LLMs in mimicking human-like reasoning and understanding.

In my initial study on this subject, I delved into commonsense reasoning, using the LLM ChatGPT to evaluate 11 benchmark datasets. Employing a questionnaire, I compared ChatGPT's responses against those of human participants, additionally questioning their evaluations of the explanations provided by the LLM. Later I broadened this inquiry by including additional open-source LLMs, such as Llama-3 by Meta and Gemma by Google, aiming for a deeper comprehension of LLMs' reasoning competencies. This endeavour focuses on contrasting the reasoning explanations generated by various LLMs for the same tasks.

Beyond commonsense reasoning, my research intends to explore syllogistic reasoning with different LLMs to analyse the field of deductive reasoning. Moreover, I'm intrigued by the potential of employing chain-of-thought prompting in reasoning tasks, a method that could

Table 1
Overview of eleven datasets for commonsense reasoning. For each dataset we report the year the dataset was published and the percentage of correct, incorrect and invalid answers of ChatGPT.

dataset	year	correct	incorrect	invalid
Story Cloze Test [15]	2017	93.33%	6,67%	0.00%
CREAK [16]	2021	86.67%	13.33%	0.00%
CODAH [17]	2019	80.00%	20.00%	0.00%
COM2SENSE [18]	2021	76.67%	23.33%	0.00%
CosmosQA [19]	2019	76.67%	23.33%	0.00%
e-CARE [20]	2022	76.67%	23.33%	0.00%
ARC [21]	2018	70.00%	30.00%	0.00%
Social IQa [22]	2019	66.67%	33.33%	0.00%
COPA [23]	2011	63.33%	3.33%	33.33%
MedMCQA [24]	2022	60.00%	40.00%	0.00%
CommonsenseQA [25]	2018	56.67%	43.33%	0.00%

further elucidate how LLMs navigate complex reasoning processes. This multifaceted approach promises to shed light on LLMs' reasoning abilities, paving the way for advancements.

To mitigate issues like hallucinations and allow better customization and scalability across various applications I aim to study RAG. I believe RAG can enhance the models' ability to provide precise, up-to-date, and contextually relevant information.

5. Preliminary Results

In this section, I present the results of my paper *Commonsense Reasoning and Explainable Artificial Intelligence Using Large Language Models* [14] presented at the European Conference on Artificial Intelligence 2023.

We analysed 11 benchmark datasets specifically curated to challenge solvers lacking commonsense knowledge. We randomly select 30 examples from each dataset. These tasks span diverse domains, including medicine, physics, and scenarios from daily life. Our evaluation of ChatGPT's capabilities using these QA benchmarks reveals a spectrum of performance outcomes. Notably, ChatGPT's weakest performance was observed on the CommonsenseQA dataset, where it achieved an accuracy of 56.67%, while its strongest performance was recorded on the Story Cloze Test, reaching an impressive accuracy of 93.33%. A detailed representation of the performance on each of the eleven datasets is shown in Table 1. Over all datasets ChatGPT answered with an accuracy of 73.33%, 77 tasks were answered incorrectly (23.33%), and we did not get a valid response for 11 QA tasks (3.33%). Not valid means that ChatGPT does not respond which answer option is correct and instead asks for further context information. We conducted an error analysis and found that there are six kinds of problems where Chat- GPT still struggles with:

1. missing context
2. comparative reasoning

3. subjective reasoning
4. slang, unofficial abbreviations, and youth language
5. social situations
6. medical domain

In our extensive questionnaire with 49 participants we found that the participants answered 73.72% of the 20 QA tasks correctly compared to ChatGPT's 90.00% on the same questions.

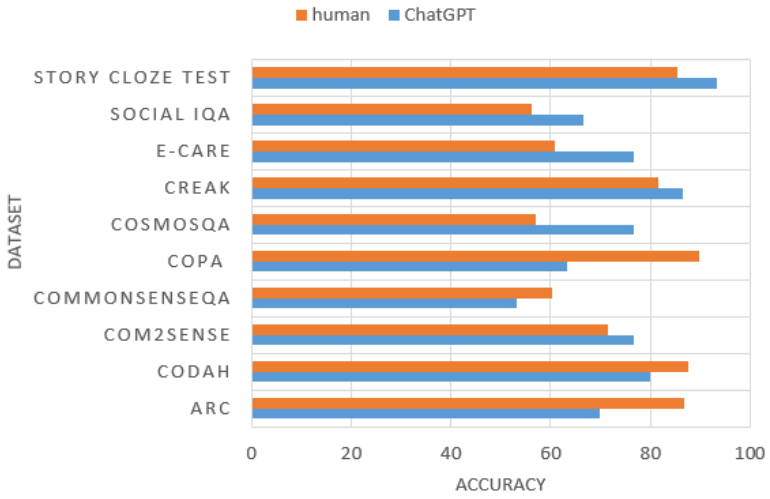

Figure 2: Comparison of accuracy of ChatGPT (blue) and our survey participants (orange) on ten different QA datasets.

Figure 2 compares ChatGPT's performance to that of surveyed participants across the datasets. ChatGPT outperformed humans on six datasets, while humans excelled in four, notably struggling with CommonsenseQA where ChatGPT also had its lowest performance. The biggest performance gap was observed in the COPA and Cosmos QA datasets, with humans outperforming ChatGPT by 26.47% in COPA and ChatGPT surpassed humans by 19.53% in Cosmos QA. Interestingly, ChatGPT showed strength in Cosmos QA, which requires contextual commonsense reasoning, despite humans significantly outperforming ChatGPT in COPA, which demands an understanding of cause and effect, and selecting the most plausible option. The findings suggest ChatGPT struggles with comparative reasoning where multiple plausible options exist, hinting that traditional explicit reasoning approaches might fare better in such scenarios. We further evaluated the explanations given by ChatGPT with the help of a questionnaire and found that explanations were mostly rated "good" or "excellent" with 67.60% and only 42 times very poor. Explanations were rated "fair" or better with 84.80%. See Figure fig:explanation for more details.

The study demonstrates that ChatGPT achieved a 73.33% overall accuracy rate on eleven QA datasets, which require commonsense reasoning for correct answers. Despite certain issues, ChatGPT managed to surpass our survey participants in six of ten datasets (excluding the MedMCQA medical dataset), suggesting that LLMs like ChatGPT are approaching near-human performance in commonsense reasoning within QA tasks. Additionally, the research also delved

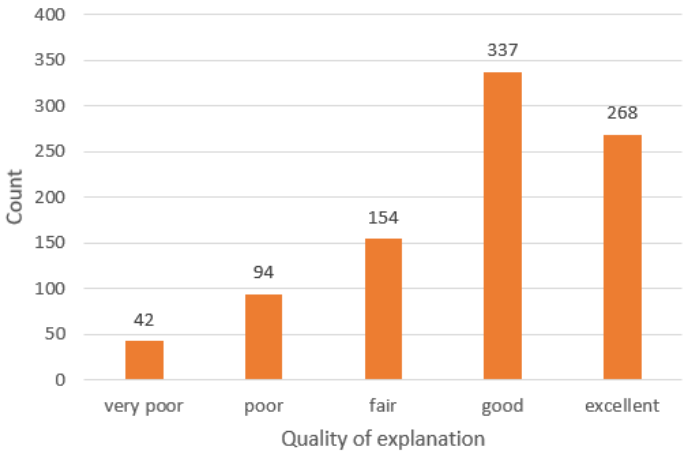

Figure 3: Participants' rating of all explanations from *very poor* to *excellent*.

into the explainability of LLMs, a critical facet in addressing the opacity of these black-box systems. According to our questionnaire, most of ChatGPT's explanations were rated "good" or "excellent", supporting our hypothesis that LLMs are capable of producing high-quality explanations.

A recent extension of this work [26], in which further LLMs have been analysed and a larger human-centred study was conducted, is currently under review.

6. Next Steps

My next steps for further research are:

1. Expand the analysis to more LLMs: Broaden the scope of the study to include a variety of LLMs to evaluate their capabilities in commonsense reasoning. This will provide a comparative understanding of different models' strengths and weaknesses in this area.
2. Investigate syllogistic reasoning abilities: Conduct in-depth analysis of the syllogistic reasoning abilities of LLMs. This involves assessing how well these models can perform logical deductions based on premises, which is a critical aspect of human-like reasoning and decision-making.
3. Assess the impact of chain-of-thought prompting: Explore how chain-of-thought prompting influences the performance of LLMs. This approach, which involves prompting models to outline their reasoning step-by-step, could enhance both the accuracy of responses and the quality of explanations provided by LLMs.
4. Evaluate the quality of LLM-generated explanations: Systematically assess the quality of explanations generated by LLMs. This could involve several criteria such as clarity, completeness, and correctness. Understanding the explanatory capabilities of LLMs is vital for their applicability in education, decision support, and other areas requiring interpretability.

5. Study the influence of RAG to receive more accurate, customized and context-aware LLM responses. The aim is to identify the strengths and limitations of RAG. The focus could be again on both the explanation quality as all as the accuracy of information.

These steps will contribute to a deeper understanding of the capabilities and limitations of LLMs, particularly in tasks requiring sophisticated reasoning and explanations.

Acknowledgments

I like to thank my supervisors Frieder Stolzenburg and Ute Schmid, for their invaluable guidance and support. Their expert advice and insightful feedback are crucial to this thesis.

References

[1] D. Gunning, D. Aha, DARPA's explainable artificial intelligence (XAI) program, AI magazine 40 (2019) 44–58. https://doi.org/10.1609/aimag.v40i2.2850.

[2] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications 10 (2019) 1096. URL: https://www.nature.com/articles/s41467-019-08987-4, 10.1038/s41467-019-08987-4.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. 10.1016/j.inffus.2019.12.012.

[4] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).

[5] S. Krause, B. H. Panchal, N. Ubhe, The evolution of learning: Assessing the transformative impact of generative ai on higher education, arXiv preprint arXiv:2404.10551 (2024).

[6] N. F. Rajani, B. McCann, C. Xiong, R. Socher, Explain yourself! leveraging language models for commonsense reasoning, arXiv preprint arXiv:1906.02361 (2019).

[7] S. Siebert, C. Schon, F. Stolzenburg, Commonsense reasoning using theorem proving and machine learning, in: Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings 3, Springer, 2019, pp. 395–413.

[8] J. Álvez, P. Lucio, G. Rigau, Adimen-sumo: Reengineering an ontology for first-order reasoning, International Journal on Semantic Web and Information Systems (IJSWIS) 8 (2012) 80–116.

[9] A. d. Garcez, K. Broda, D. M. Gabbay, Symbolic knowledge extraction from trained neural networks: A sound approach, Artificial Intelligence 125 (2001) 155–207.

[10] L. Huang, R. Le Bras, C. Bhagavatula, Y. Choi, Cosmos QA: Machine reading comprehension with contextual commonsense reasoning, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2391–2401. URL: https://aclanthology.org/D19-1243. doi:10.18653/v1/D19-1243.

[11] L. Pan, A. Albalak, X. Wang, W. Y. Wang, Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning, arXiv preprint arXiv:2305.12295 (2023).

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[13] D. Paul, M. Ismayilzada, M. Peyrard, B. Borges, A. Bosselut, R. West, B. Faltings, Refiner: Reasoning feedback on intermediate representations, 2024. arXiv:2304.01904.

[14] S. Krause, F. Stolzenburg, Commonsense reasoning and explainable artificial intelligence using large language models, in: European Conference on Artificial Intelligence, Springer, 2023, pp. 302–319.

[15] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, J. Allen, LSDSem 2017 shared task: The Story Cloze Test, in: Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, 2017, pp. 46–51. URL: https://aclanthology.org/W17-0906.pdf.

[16] Y. Onoe, M. J. Q. Zhang, E. Choi, G. Durrett, CREAK: A dataset for commonsense reasoning over entity knowledge, 2021. URL: https://arxiv.org/pdf/2109.01653.

[17] M. Chen, M. D'arcy, A. Liu, J. Fernandez, D. Downey, CODAH: An adversarially-authored question answering dataset for common sense, in: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, 2019, pp. 63–69. URL: https://www.jaredfern.com/publication/codah/.

[18] S. Singh, N. Wen, Y. Hou, P. Alipoormolabashi, T.-L. Wu, X. Ma, N. Peng, COM2SENSE: A commonsense reasoning benchmark with complementary sentences, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, 2021, pp. 883–898. URL: https://aclanthology.org/2021.findings-acl.78.

[19] L. Huang, R. Le Bras, C. Bhagavatula, Y. Choi, Cosmos QA: Machine reading comprehension with contextual commonsense reasoning, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 2391–2401. URL: https://aclanthology.org/D19-1243/.

[20] L. Du, X. Ding, K. Xiong, T. Liu, B. Qin, e-CARE: a new dataset for exploring explainable causal reasoning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 432–446. URL: https://aclanthology.org/2022.acl-long.33/.

[21] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, CoRR – Computing Research Repository abs/1803.05457, Cornell University Library, 2018. URL: https://arxiv.org/abs/1803.05457.

[22] M. Sap, H. Rashkin, D. Chen, R. LeBras, Y. Choi, Social IQa: Commonsense reasoning about social interactions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 4463–4473. URL: https://aclanthology.org/D19-1454/.

[23] M. Roemmele, C. A. Bejan, A. S. Gordon, Choice of plausible alternatives: An evaluation of commonsense causal reasoning., in: AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011, pp. 90–95. URL: https://aaai.org/papers/02418-choice-of-plausible-alternatives-an-evaluation-of-commonsense-causal-reasoning/.

[24] A. Pal, L. K. Umapathi, M. Sankarasubbu, MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering, ACM Conference on Health (2022). URL: https://arxiv.org/pdf/2203.14371.

[25] A. Talmor, J. Herzig, N. Lourie, J. Berant, CommonsenseQA: A question answering challenge targeting commonsense knowledge, 2018. URL: https://arxiv.org/pdf/1811.00937.

[26] S. Krause, F. Stolzenburg, From data to commonsense reasoning: The use of large language models for explainable ai, arXiv preprint arXiv:2407.03778 (2024).