Enhancing the Explainability of Large Language Models

Xizhi Xiao
Yuanpei College
Peking University
xizhixiao.pku@gmail.com

Abstract

Deep neural networks, including large language models, have faced criticism for their limited explainability. In this essay, we undertake a review of recent research aimed at augmenting the explainability of these models. We posit two challenges that demand attention in future investigations: 1) Providing explanations grounded in factual knowledge, and 2) Incorporating contextual knowledge for enhanced explainability. We contend that addressing these two challenges is pivotal to improving the overall explainability of large language models.

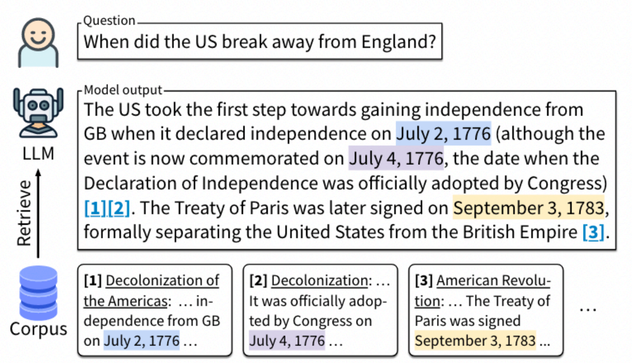
1 Introduction

Humans inherently possess a proclivity for elucidating the intricacies of the world surrounding them. Their curiosity extends beyond a mere understanding of how the world functions; they are equally eager to unravel the reasons behind its functioning. Conversely, for effective communication and validation, humans find it imperative to explicate their actions and decisions to others [2]. With the advent of artificial intelligence, an increasing number of decisions are delegated to machines. These models, reliant on copious parameters and extensive training datasets, undergo end-to-end training, rendering the intermediate processes akin to a black box. This opacity significantly diminishes the explainability of the model, emerging as a substantial impediment to the widespread application of artificial intelligence across diverse domains. Notably, in fields like medicine, the inscrutability of machine learning models poses challenges, hindering doctors from placing trust in the diagnostic outcomes generated by these models.

In recent years, a multitude of researchers has dedicated efforts to enhance the explainability of deep neural networks. A notable approach employed in large language models is Chain-of-Thought Prompting [8]. This method involves guiding the Chatbot through a step-by-step thought process, resulting in improved performance in solving mathematical problems and addressing issues related to abstract reasoning. Delving deeper into the efficacy of Chain-of-Thought prompting, Prystawski and Goodman [6] conducted a comprehensive investigation. Their findings revealed that the success of thinking step by step is rooted in the local statistical nature of the training data. Specifically, when the training data consists of overlapping local clusters of variables, the training conditions enhance relationships between these local variables, thereby facilitating accurate local inferences. In addition to directly outputting the thinking process of large language models to enhance explainability, Karataş and Cutright [5] demonstrated that contemplating notions of God positively influences individuals' acceptance of recommendations from AI systems. This underscores the importance of considering human beliefs and values in the design of AI systems, a factor that proves beneficial to the overall explainability of the model.

2 Explain with factual knowledge

Large language models have faced criticism for their propensity for hallucination [1]. In the process of generating responses to human input, these models exhibit a tendency to produce texts that contain fictitious information, including fabricated examples or citations [4], and facts that lack accuracy. Furthermore, even when incorporating the correct reference, the generated texts may lack consistency with the content found in that reference. These issues pose a significant challenge, making it arduous for users to place trust in the answers provided by large language models.



When did the US break away from England?

The US took the first step towards gaining independence from GB when it declared independence on July 2, 1776 (although the event is now commemorated on July 4, 1776, the date when the Declaration of Independence was officially adopted by Congress) [1][2]. The Treaty of Paris was later signed on September 3, 1783, formally separating the United States from the British Empire [3].

[1] Decolonization of the Americas: … independence from GB on July 2, 1776 …

[2] Decolonization: … It was officially adopted by Congress on July 4, 1776 …

[3] American Revolution: … The Treaty of Paris was signed September 3, 1783 …

Figure 1: Given questions, the system generates answers with citations as supporting evidence. (Adopted from [3])

To address this challenging problem, Gao et al. [3] focus on the question-answering task, a prominent aspect of natural language generation. In their efforts to enhance the correctness and verifiability of generated texts, they empower large language models to produce output with citations as supporting evidence. As depicted in Fig. 1, the system, when presented with a question, retrieves pertinent citations from the corpus and generates an answer with citations serving as supporting evidence. The evaluation of responses encompasses three dimensions: fluency, correctness, and citation quality. Through their experiments, Gao et al. [3] illustrate that even state-of-the-art large language models, such as GPT-4 and LLaMA, exhibit significant potential for improvement in citation quality and correctness.

3 Explain with contextual knowledge

Humans possess the ability to employ contextual knowledge to facilitate communication and explanation, emphasizing the importance of a clear understanding of shared common ground for effective interaction [7]. To communicate efficiently, individuals must be aware of what others already know, allowing them to tailor their expressions accordingly. For instance, when elucidating the cause of a disease to a child, a doctor employs familiar language instead of professional terminology. This principle extends to large language models, which must harness contextual knowledge to effectively convey their thoughts to humans. Achieving this requires the application of techniques such as value alignment and adherence to communicative conventions.

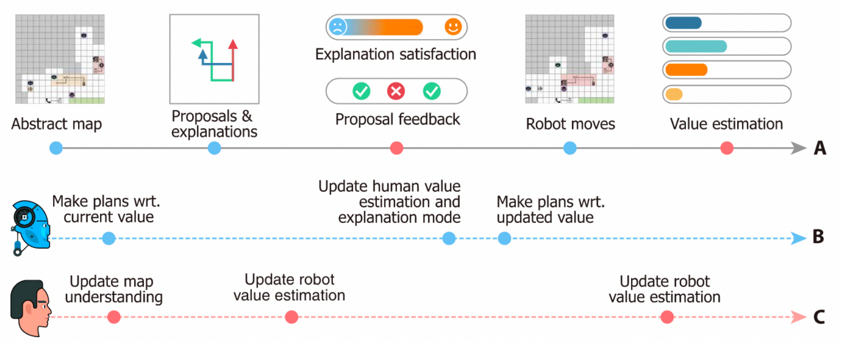

Abstract map
Proposals & explanations
Explanation satisfaction
Proposal feedback
Robot moves
Value estimation
A
Make plans wrt. current value
Update human value estimation and explanation mode
Make plans wrt. updated value
B
Update map understanding
Update robot value estimation
Update robot value estimation
C

Figure 2: The process of bidirectional human-robot value alignment in a certain task. (Adopted from [9])

Addressing the challenge of value alignment through communication, Yuan et al. [9] propose a bidirectional human-robot value alignment framework. In this framework, the robot deduces human intentions from feedback and explains its decision-making process to users. As illustrated in Fig. 2, the system integrates a module to infer human values related to multiple goals, facilitating one direction of value alignment. In the opposite direction, the system generates optimal explanations based on the user's mental states using graphical models. Drawing inspiration from this model, we posit that bidirectional value alignment is crucial for large language models when communicating with human users.

4 Conlusion

In conclusion, the lack of explainability in large language models presents a significant obstacle to their widespread adoption. We have delineated two pivotal challenges for enhancing model explainability: explaining with factual knowledge and explaining with contextual knowledge. Future research must continue to address these challenges and explore innovative approaches to improve the overall explainability of large language models.

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023. 1

[2] Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967. 1

[3] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023. 2

[4] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 2

[5] Mustafa Karataş and Keisha M Cutright. Thinking about god increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences*, 120(33): e2218961120, 2023. 1

[6] Ben Prystawski and Noah D Goodman. Why think step-by-step? reasoning emerges from the locality of experience. *arXiv preprint arXiv:2304.03843*, 2023. 1

[7] Michael Tomasello. *Origins of human communication*. MIT press, 2010. 2

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 1

[9] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183, 2022. 2