LLM Reliability and CBR: How Case Based Reasoning Can Improve the Performance of Large Language Models

Kaitlynne Wilkerson[1,*]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47408, USA

Abstract

Large Language Models (LLMs) are known to make factual errors and hallucinate. This project overview discusses current and future research methods of improving the accuracy, interpretability and explainability of LLMs leveraging knowledge obtained by Case Based Reasoning.

Keywords

Large Language Models, ChatGPT, Case Based Reasoning, Trusted AI, Llama

1. Introduction

Over the last decade, AI technology has become increasingly integrated into devices and services, which means that AI failures can have a wider and potentially more damaging impact. These failures may take the form of unpredictable, confusing, disruptive, offensive, or even dangerous behavior and are made possible due to the black box nature of many AI architectures, difficulty in keeping training data up-to-date, and the possibility of performing under uncertainty [1]. Frequent instances of failure, or even infrequent instances with severe consequences, can damage human trust in AI technology and may even lead to an aversion to it [1, 2]. Although some measures have been taken to regulate AI technology (e.g., EU GDPR), trust and reliability remains a critical area of AI research. The introduction of ChatGPT turned both the public and research community's conversation towards the capabilities of Generative AI (GenAI), particularly Large Language Models (LLMs). While LLMs have demonstrated an impressive ability to generate human-like language, they also encapsulate most of the risks discussed above. Given how enraptured the world has been by ChatGPT and how many have rushed to incorporate LLMs into existing systems, LLMs present, in my opinion, one of the greatest threats to public trust in AI systems to date.

What Does it Mean to Trust an AI System?: Since human-to-computer interaction contains many of the same sociological underpinnings of human-to-human interaction, interpersonal trust can be used to understand the core components for improving human trust in AI responses [3, 4]. Confidence is the key aspect to interpersonal trust, specifically confidence in knowing that an agent can perform a certain task at a level that is comfortable to the risk one is taking. When AI systems present the wrong solution to a human user, it negatively impacts their trust assessment of the system [5], which indicates that the correct solution is an expectation of users. Furthermore, the ability to explain how a solution was derived is considered a crucial design decision and systems lacking explanatory capabilities tend to be viewed more negatively [4, 2, 1, 6]. From this, it can be argued it is critical for AI systems to produce the correct solution and be able to explain how they arrived at this decision. The **primary objectives** of this project will be to improve the accuracy of LLMs, make their reasoning process more interpretable and explaining that process in an easy-to-understand fashion.

Improving Trust in Large Language Models: LLMs have an immense amount of information encoded into the network that may be exploited for more complex tasks but limitations, such as hallucinations and factual errors, introduce risk to user trust [7, 8]. These limitations stem from the model's lossy knowledge encoding, which can lead to knowledge generalizations and distortions [9], and asking the

ICCBR DC'24: Doctoral Consortium at ICCBR2024, July 1, 2024, Mérida, Mexico

*Corresponding author.

kwilker@iu.edu (K. Wilkerson)

LLM for knowledge beyond the data it was trained on [10]. To mitigate associated risks, it is essential to focus on improving the accuracy and plausibility of the responses that LLMs provide. Many approaches for improving LLMs have proven to be computationally, financially and temporally expensive [7, 9]. This makes plug-and-play (PnP) approaches for improving LLM performance, such as external knowledge integration, particularly attractive [9]. One such method is Retrieval Augmented Generation (RAG), which retrieves chunks of vectorized text based on similarity to a query [10]. This knowledge can then be integrated into the LLM prompt and has been shown to improve LLM accuracy [11].

Case Based Reasoning (CBR) and RAG both work to provide similar information to a problem query, but a difference exists primarily in the type of information retrieved. RAG returns knowledge statements about the domain that the LLM operates in and can be thought of as analogous to human semantic memory [12]. Case Based Reasoning, however, provides concrete episodes (i.e., cases) regarding the domain task and can be considered akin to human episodic memory [12]. Explanations built from cases have shown to be more convincing than explanations built from domain-based rules and the CBR process mimics human reasoning methods [13, 14]. Along with the inherent transparentness of CBR methods and their established use with black box methods to improve explainability [13, 15], cases and CBR possess immense potential to improve the interpretability and explainability of LLM responses and, in turn, human trust. However, as the effects of episodic knowledge on LLM accuracy have not been well documented, it is necessary to establish cases as a method for improving accuracy.

Research Impacts: This research will provide much needed knowledge on the integration of CBR and LLMs as well as establish cases as an attractive knowledge source for LLM improvement.

2. Research Plan

2.1. Research Objectives

This project seeks to improve the accuracy, interpretability, and explainability of LLMs. There will most likely be three stages of research. Stage 1, which is currently underway, will work to articulate methods for case-augmented generation that improve performance over LLM baselines and provide cases as an alternative to RAG. Stage 2 will attempt to extend the work of Stage 1 by focusing on conditions where case related problems may introduce confusion into the LLM's reasoning process. Finally, Stage 3 will most likely be focused on user's reception of LLM explanations generated under conditions tested in stages 1 and 2. Further information on the stages will be provided in the Future Work section.

2.2. Approach / Methodology

Previous Work: A human subject study on CBR generated explanations and trust was completed and will be used to better understand how to utilize CBR for trusted explanation development [5]. Another paper, exploring the ability for cases to improve LLM accuracy, has been accepted for publication to the ICCBR 2024 Main Conference [16].

LLMs Used: ChatGPT 3.5[1] and Llama 2 70B Chat [17] were used in the exploratory experiments discussed in [16]. ChatGPT was chosen to illustrate the abilities of current commercial state-of-the-art models and Llama 2 was selected for replicability purposes. [16] contains more detailed information about model deployment. Future experiments will likely continue to build on the initial results obtained with these models. Other open source LLMs may also be tested in the future.

Case Base Development: The experiment in [16] used a triage classification dataset obtained on Kaggle.[2] The dataset was pre-processed to remove any instances containing missing values. From the remaining data, the training and test sets were a randomly selected subset containing representation from each class.

The dataset used in this experiment was released in 2019, which means that either test LLM may have been trained on the data. As a result, we would expect to see the LLM baseline to have better

[1] https://openai.com/blog/chatgpt

[2] https://www.kaggle.com/datasets/ilkeryildiz/emergency-service-triage-application

performance and that the impact of cases would be dampened. We did not observe this phenomenon in the results, and while we do not know if either LLM was trained on the dataset, the results seem to indicate that a benefit still exists to providing cases. Since the results from this set of experiments were only meant to be exploratory, future experiments will likely seek to obtain data that is outside of the public domain or was introduced after the last set of model training.

k-NN Retrieval: k-NN retrieval was used as a performance baseline and used to select cases to present to the LLM. Feature weights were selected via hill climbing and non-numerical values in the dataset were assessed by: assigning a distance of 1 for non-matching categorical data and using cosine similarity on vectorized text strings to assess semantic similarity. Although the datasets for future experiments are currently undecided, all data will be put through a similar process and k-NN will continue to be used in the same manner.

Prompt Construction: The wording of each prompt type was decided through an extensive round of pre-testing to understand how LLMs responded to different phrasings of the same task. While the details of that pre-testing are omitted from [16] due to space, the best performing phrases were used in the final experiment. Future experiments will likely reuse or be derivations of these prompts but change any domain relevant details.

3. Progress Summary

As of the time of publication, a literature review has been conducted on topics related to the methods and concerns outlined so far, a human subjects study examining the impact of cases and their presentation on user trust [5] has been published and an exploratory study on the benefit of cases on LLM accuracy is up for publication at ICCBR 2024 [16]. These efforts have allowed for the development of a tentative schedule for a research program examining the benefit of case knowledge and CBR as a methodology for improving LLMs. The remainder of this section will focus on the results gleaned from the completed work.

ICCBR 2023: In [5], we performed a human subjects study that tested the impact that various case knowledge formulations used in AI generated explanations had on user trust. The results demonstrated that providing the nearest neighbor along with explicit statements of difference in tabular form between the problem case and nearest neighbor elicited the highest user scores on trust. Providing only the nearest neighbor and the nearest neighbor plus explicit difference statements in textual form also performed well for user trust. We also tested whether the AI providing an intentionally incorrect solution and the similarity level between the problem and solution cases had an implicit effect on user scoring. Our tests concluded that, while users were not explicitly aware of these conditions, intentionally incorrect solutions and lower similarity levels between problem and solution resulted in lower scores. Since this project aims to improve the explainability of LLMs as method of improving user trust, these results justify the usage of the nearest neighbor in explanations while underpinning the importance of similarity assessment and solution correctness in model interpretability.

ICCBR 2024: Using a triage classification task, we conducted a simple experiment comparing the baseline accuracy of ChatGPT and Llama 2 against the accuracy of each LLM when it has access to cases and is either prompted to perform a sort of implicit CBR (ICBR)—to solve the new problem using a provided case—or a more explicit CBR (ECBR), in which cases are provided and the LLM is prompted to perform the steps of CBR, specifically similarity assessment and adaptation [16]. The accuracy tests demonstrated a clear pattern of cases improving the performance of LLMs over direct solution baselines and established ICBR provided with the Nearest Neighbor (1NN) as the best performing prompt type across models. Additionally, ICBR with the top 2 Nearest Neighbors (2NN) performed well on Llama 2 and ECBR 2NN performed well on ChatGPT. The difference in performance of ICBR 2NN and ECBR 2NN on each model may be due in part to the relatively low adaptation rates of Llama 2. ChatGPT and Llama 2 were found to perform similarity assessment at roughly the same rates, but ChatGPT performed more adaptation. Because Llama 2 does not do adaptation as often, this likely hurt the accuracy of the ECBR 2NN prompt on Llama 2. These results suggest that the individual capabilities of

LLMs may affect the impact that CBR can have on accuracy. In conjunction with the results from [5], the similarity assessment capabilities of LLMs and the correctness of the response are critical aspects of LLM performance with respect to user trust.

4. Conclusion and Future Work

This section will provide more details regarding the stages discussed in the Research Objectives section.

Stage 1: Although definitive plans have not yet been set, there are currently three questions that may be beneficial for this stage:

1. Can case integration into prompts improve LLM performance over a baseline no-knowledge prompt? Are these results generalizable over same domain tasks and different domains?
2. Can case integration into prompts perform equivalently to RAG?
3. Are there ways to improve LLM capabilities with respect to Case Based Reasoning?

[16] begins to explore question 1. These methods will also be tested on multiple different tasks and domains and along with other open source LLMs, such as Llama 3. Results from these tests will help to further elucidate the impact that cases and CBR can have on LLM accuracy. Given that one of the end goals of this stage is to be able to compare case-augmented knowledge integration methods against RAG, it will be necessary to test how RAG performs in these domains. Finally, based on the initial results from [16], it appears that differences may exist in LLM ability to utilize cases to their fullest extent. It may be fruitful to explore whether and how these differences may be accounted for during interaction with the goal of making model performance generalizable. This may be tested by providing examples of ideal behavior to the system during the prompting phase or developing a secondary case base that can be used to query examples of ideal behavior when needed. While the final project may not follow these investigative questions exactly, it is indicative of the types of experiments that stage one will consist of.

Stage 2: This stage will focus on conditions where imperfections in integrated data can introduce confusion into the LLM's reasoning process, such as partial information in a problem or prior case, or a lack of representation of certain classes. This stage will partially be focused on the observation of LLM behavior under these conditions and how errors in reasoning may appear as well as developing methods for addressing behavioral problems.

Partial Information: Data gleamed from real world scenarios are likely to contain instances where information is missing, which may translate to missing values in the problem case or cases in the case base. Because LLMs have broad knowledge but little depth of knowledge, it is highly likely that LLMs may try to fill in knowledge gaps with generalizations that could confuse or introduce harm. One way to address this might be to introduce secondary or tertiary cases to provide more domain specific information or to attempt to leverage CBR processes to adapt network generalizations with case information. In this application, cases may be more useful for providing problem specific information or for guiding specification of LLM responses.

Lack of Class Representation: This type of problem is fairly similar to partial knowledge situations, except instead of missing information existing at the case level, the missing or lack of knowledge sits at the classification level. If the case base does not contain an example for a possible classification or the only examples of a classification are not similar enough, case information alone may not be enough to solve the problem. This is where the knowledge embedded in the LLM may be particularly useful; it may be able to generalize very specific knowledge that cases contain in order to differentiate class boundaries.

Stage 3: This stage will focus on how human users evaluate their trust in LLM generated explanations under the conditions and using the methods of stages 1 and 2. While it may not be possible to test every single method and condition described in the previous sections, core aspects to be tested include the differences in trust between explanations generated using case data, RAG data and no-knowledge and the impact on trust of LLM behavior when missing data values or lack of class representation exist.

Acknowledgments

This work was funded by the US Department of Defense (Contract W52P1J2093009). This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

[1] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al., Guidelines for human-AI interaction, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–13.

[2] S. S. Sundar, Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII), Journal of Computer-Mediated Communication 25 (2020) 74–88.

[3] M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management, Big Data & Society 5 (2018) 2053951718756684.

[4] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 624–635.

[5] L. Gates, D. Leake, K. Wilkerson, Cases are king: A user study of case presentation to explain CBR decisions, in: International Conference on Case-Based Reasoning, Springer, 2023, pp. 153–168.

[6] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, A. Dhurandhar, Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable AI, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 10, 2022, pp. 147–159.

[7] K. Hammond, D. Leake, Large language models need symbolic AI, in: Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, volume 3432, 2023, pp. 204–209.

[8] B. Paranjape, J. Michael, M. Ghazvininejad, L. Zettlemoyer, H. Hajishirzi, Prompting contrastive explanations for commonsense reasoning tasks, arXiv preprint arXiv:2106.06823 (2021).

[9] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, et al., Check your facts and try again: Improving large language models with external knowledge and automated feedback, arXiv preprint arXiv:2302.12813 (2023).

[10] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).

[11] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. L. Bras, Y. Choi, H. Hajishirzi, Generated knowledge prompting for commonsense reasoning, arXiv preprint arXiv:2110.08387 (2021).

[12] E. Tulving, et al., Episodic and semantic memory, Organization of memory 1 (1972) 1.

[13] D. B. Leake, CBR in context: The present and future, Case-based reasoning: Experiences, lessons, and future directions (1996) 3–30.

[14] P. Cunningham, D. Doyle, J. Loughrey, An evaluation of the usefulness of case-based explanation, in: International conference on case-based reasoning, Springer, 2003, pp. 122–130.

[15] I. Watson, A case-based persistent memory for a large language model, arXiv preprint arXiv:2310.08842 (2023).

[16] K. Wilkerson, D. Leake, On implementing case-based reasoning with large language models, in: International Conference on Case-Based Reasoning, Springer, in press.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).