Title: LLM Explainability Explored through College Recommendations

Author(s): Lindey Carl, Mizanur Rahman, Michael Brewer

Abstract:

Large Language Models (LLMs) are being used, successfully, in a wide variety of problems. With the growth in popularity of these models, it becomes more critical to explain why certain results are generated. As engineers, doctors, lawyers, and others begin to rely on systems utilizing these models, it becomes imperative to understand how inputs result in output, and the complexity and size of LLMs make these explanations challenging.

In this project, we will utilize LLMs to provide College Recommendations, and this problem will serve as an example of a Critical Decision-Making application of LLMs. We systematically

1. Construct a synthetic data set of students with a set of features including age, gender, race, area of interest, colleges of interest with projected costs, standardized test scores, and possible other features. Note that this dataset will not include a ground-truth result, as our interest is not specifically what recommendation the LLM provides but what contributed to the recommendation being selected.
2. On a per-student basis, we will systematically create prompts using zero-shot, n-shot, and chain-of-thought techniques within OpenAI's ChatGPT.
3. We will then utilize perturbation techniques to assess saliency values, the features built into our prompts and analyze the resulting data.
4. Additionally, we utilize a model with open-source weights to evaluate a gradient-based explanation for the zero-shot results.

Keywords: Large Language Models, LLMs, Transformers, Explainability, Interpretability, Ethical AI, XAI, Recommender

References:

[1] W. Ding, M. Abdel-Basset, H. Hawas and A. M. Ali, "Explainability of artificial intelligence methods, applications," Information Sciences, vol. 615, pp. 238-292, 2022.

[2] C. Rudin, "Stop explaining black box machine learning," Nature Machine Intelligence, vol. 1.5, pp. 206-215, 2019.

[3] R. Caruana, y. G. J. Lou, P. Koch, M. Strum and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sidney, NSW, Australia, 2015.

[4] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin and M. Du., "Explainability for Large Language Models: A Survey," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 2, pp. 1-38, 2023.

[5] S. Wu and e. al, "Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions," arXiv preprint, vol. 2307, no. 13339, 2023.

[6] A. a. Y. A. Madaan, "Text and patterns: For effective chain of thought, it takes two to tango," arXiv preprint arXiv:2209.07686, 2022.

[7] Amazon Web Services, "Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions," 2022. [Online]. Available: https://aws.amazon.com/ecs/. [Accessed 5 November 2022].

[8] B. T. B and e. al, "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020.

[9] C. Singh, A. Askari and R. e. a. Caruana, "Augmenting interpretable models with large language models during training.," Nature Communications, vol. 14, no. 7913, 2023.