World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:02, 2024

Large Language Model Powered Chatbots Need End-to-End Benchmarks

Authors : Debarag Banerjee, Pooja Singh, Arjun Avadhanam, Saksham Srivastava

Abstract : Autonomous conversational agents, i.e., chatbots, are becoming an increasingly common mechanism for enterprises to provide support to customers and partners. In order to rate chatbots, especially ones powered by Generative AI tools like Large Language Models (LLMs), we need to be able to accurately assess their performance. This is where chatbot benchmarking becomes important. In this paper, authors propose the use of a benchmark that they call the E2E (End to End) benchmark and show how the E2E benchmark can be used to evaluate the accuracy and usefulness of the answers provided by chatbots, especially ones powered by LLMs. The authors evaluate an example chatbot at different levels of sophistication based on both our E2E benchmark as well as other available metrics commonly used in the state of the art and observe that the proposed benchmark shows better results compared to others. In addition, while some metrics proved to be unpredictable, the metric associated with the E2E benchmark, which uses cosine similarity, performed well in evaluating chatbots. The performance of our best models shows that there are several benefits of using the cosine similarity score as a metric in the E2E benchmark.

Keywords : chatbot benchmarking, end-to-end (E2E) benchmarking, large language model, user centric evaluation.
Conference Title : ICEMNLP 2024 : International Conference on Empirical Methods in Natural Language Processing
Conference Location : Bangkok, Thailand
Conference Dates : February 01-02, 2024