Evaluating Performance Metrics in Bias Mitigation for Generated Patient Vignettes using Large Language Models

Shaurya Kumar, Chinmay Agrawal

Recent advancements in Large Language Models (LLMs) have introduced new possibilities for automating medical vignette generation, a process that could significantly aid in patient diagnosis. However, concerns about bias in the outputs of these models necessitate a comprehensive evaluation of their performance and fairness. Our study leverages a sophisticated LLM pipeline to generate medical vignettes from a dataset curated from clinical guidelines, biomedical literature, and clinical trials. The LLM pipeline involves multiple stages: a context retriever that loads and embeds clinical guidelines, biomedical literature, and clinical trials; a prompt generator that creates specific prompts based on user queries; a strong LLM (e.g., GPT-4) that generates vignettes; and a post-processing stage that refines the outputs for benchmarking. This workflow ensures that the generated vignettes are contextually relevant and tailored to the specified medical conditions.To explain the motivations for using these specific metrics, we delve into the statistical and model-based scoring methods. ROUGE and BLEU scores measure n-gram overlaps, while GPTScore and Semantic Entropy assess semantic similarity and fluency. Our dataset comprises clinical guidelines and biomedical literature, and we detail the sources and examples used in our analysis. By performing rigorous tests, we present the numerical outcomes of these metrics, demonstrating their significance in evaluating the quality of generated patient vignettes. In conclusion, our research emphasizes the importance of selecting appropriate evaluation metrics to ensure the accuracy and reliability of patient vignettes generated by LLMs. These findings contribute to improving the effectiveness of LLMs in clinical settings, ultimately aiding in unbiased and accurate patient diagnosis.