Elton Cardoso do Nascimento - 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "RAGAs: Automated Evaluation of Retrieval Augmented Generation" (Es et al.)

O artigo apresenta uma nova técnica para avaliar sistemas de RAG. Devido ao fato da performance desses sistemas ser dependente de seus componentes, a possível não existência de respostas de referência e o possível não acesso ao LLM sendo utilizado, é complexo criar uma forma para avaliá-los. O trabalho sugere uma avaliação automatizada utilizando uma LLM, baseada em 3 métricas:

Fidelidade avalia se a resposta gerada é baseada no contexto, evitando alucinações. Ela funciona extraindo um conjunto de sentenças a partir da resposta, e verificando se estas sentenças podem ser obtidas através do contexto.

Relevância da resposta avalia se a resposta gerada responde a pergunta original. Para isso, é gerado um conjunto de perguntas a partir da resposta, e calculado a similaridade de cosseno entre estas perguntas e a pergunta original.

Relevância do contexto avalia se o contexto não contém informação irrelevante. A LLM aqui tem como tarefa extrair sentenças do contexto relevantes para a resposta, sendo avaliado a proporção entre sentenças relevantes e totais no contexto.

Para análise das métricas criadas, o dataset "WikiEval" foi criado, a partir de páginas da Wikipédia com auxílio do ChatGPT, que contém um conjunto de (questão, contexto, resposta) anotados por humanos. As anotações dos humanos são comparadas com métricas e outras técnicas, concluindo que o RAGAs possui a maior concordância com os avaliadores. Em ordem de melhor concordância estão fidelidade, relevância da resposta e relevância do contexto.