Destaques da leitura de “**RAGAS: Automated Evaluation of Retrieval Augmented Generation**”

Nome: Caio Petrucci dos Santos Rosa **RA:** 248245

Os destaques da leitura do artigo foram os seguintes:

- O texto propõe um *framework* de avaliação de sistemas RAG baseados em LLM que seja capaz de sistematicamente avaliar a performance e a qualidade das respostas desses sistemas sem considerar referências como textos ou respostas esperadas;

- Para avaliar sistemas RAG sem referências anotadas, o RAGAS considera 3 aspectos de uma resposta gerada para um pergunta: *faithfulness*, que indica se a resposta está fundamentada no contexto recuperado e dado ao LLM; *answer relevance*, que mostra se a resposta aborda a pergunta feita; e *context relevance*, que indica se o contexto recuperado é relevante dada a pergunta feita, ou seja, se o contexto recuperado possui informações suficientes para que uma resposta possa ser gerada;

- Para avaliar cada uma dos aspectos do RAGAS, foi utilizado o **GPT-3.5-Turbo** (16 mil tokens de contexto) e o modelo de *embeddings* **text-embedding-ada-002**, ou seja, foram utilizados outros modelos de linguagem nesse processo. Essa abordagem foi avaliada em um *dataset* construído pelo autores, chamado WikiEval, e a acurácia do RAGAS foi satisfatória quando comparada à anotação de humanos, principalmente para o aspecto *faithfulness*;

- Fiquei me perguntando se um *fine-tuning* de modelos de linguagem e de *embeddings*, como os utilizados para avaliação, para exatamente o tipo de tarefa que eles incorporaram no RAGAS não possibilitaria um sistema ainda mais robusto de avaliação de sistemas RAG, dado que diversos textos que lemos ao longo da disciplina explicitaram uma melhora de performance grande após *fine-tuning*, como foi o caso do trabalho original do ReAct.