Destaques da leitura de “**Retrieval-Augmented Generation for Large Language Models: A Survey**”

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

Os destaques da leitura do artigo foram os seguintes:

- O artigo faz um levantamento de trabalhos envolvendo Retrieval-Augmentation Generation (RAG), destacando as principais vantagens que esse processo traz a sistemas baseados em LLMs, como a expansão do conhecimento para além dos dados de treinamento do modelo, a redução de alucinações e a capacidade de rastrear fontes e referências. Também são apresentados 3 paradigmas de RAG, com fluxos de execução cada vez mais complexos: “Naive RAG”, “Advanced RAG” e “Modular RAG”;

- Os autores também discutem diversos desafios como escolher técnicas de processamento do texto para indexação e de recuperação, considerando a importância de recuperar informações relevantes mas não incluir informações irrelevantes neste processo. Além disso, também apontam o papel crítico da construção de bons prompts, incluindo as informações mais relevantes nas extremidades (começo ou fim do prompt);

- Achei interessante a discussão sobre o papel do RAG dado que os LLMs mais modernos estão suportando contextos cada vez maiores, que estão se tornando capazes de aceitarem documentos inteiros. Acho importante ressaltar o papel crítico que RAG desempenha na integração de sistemas mais complexos de recuperação de dados, possivelmente de diversas bases de dados, e outros fluxos, como ocorre no paradigma “Modular RAG”, dado a flexibilidade deste processo. Os autores também ressaltam um ponto interessante sobre RAG: a geração baseada em um contexto muito longo ainda é uma caixa-preta;

- Considerando o item acima, também fiquei refletindo que quanto melhor as técnicas de RAG, possivelmente se torna mais viável a integração de SLMs em sistemas que dependem inteiramente de modelos maiores com muito conhecimento interno;

- Achei interessante a discussão sobre diferentes formas de realizar a indexação dos dados, como a “structural index”, dado que existem dados complexos de serem pré-processados e recuperados como tabelas, ainda mais tabelas grandes.