Nome: Fabio Grassiotto RA: 890441
IA024 - Aula 8_9
Leitura do Artigo “Retrieval-Augmented Generation for Large Language Models A Survey”

Principais Contribuições do Artigo:

Este é um artigo de revisão sistemática sobre o assunto de RAG, *Retrieval-Augmented Generation,* que consiste em técnicas para o uso de bases de conhecimento externo para a alimentação de modelos grandes de linguagem visando melhorias na qualidade das respostas obtidas.

As principais contribuições que posso ressaltar deste artigo são:

- Apresentação de uma “árvore” tecnológica da pesquisa na área de RAG com uma revisão histórica da evolução das técnicas apresentadas.
- A divisão das técnicas apresentadas em três categorias:
    - *Naive RAG*
    - *Advanced RAG*
    - *Modular RAG*

  Com uma explicação dos principais módulos e funcionalidades das técnicas de RAG em cada uma das categorias.
- A identificação e sumarização dos principais desafios nos campos de Busca de Documentos (*Retrieval),* Geração de respostas dos modelos de linguagem, processos de acréscimo em RAG (*Augmentation)* e os sistemas de avaliação para as tarefas downstream.
- Uma discussão dos desafios que ainda existem no processo de RAG descrevendo alguns caminhos possíveis para a pesquisa na área, incluindo o aumento do tamanho do contexto em documentos externos, a robustez do processo de RAG, propostas híbridas (como a combinação de RAG com o processo de *fine-tuning),* aumento de escala, aspectos de produção e RAG multi-modal.