Comentário crítico: *Retrieval-Augmented Generation for Large Language Models: A Survey*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo discute a técnica de RAG (*Retrieval-Augmented Generation*), que combina o uso de LLMs com a busca externa de documentos antes ou durante a geração de respostas, visando melhorar a qualidade dos resultados. Um sistema RAG é composto por dois elementos principais: o *retriever*, responsável por buscar informações relevantes em uma coleção de documentos externa, e o *generator*, que recebe os documentos coletados juntamente com o prompt do usuário, utilizando um LLM para gerar as respostas.

O RAG ganha destaque no contexto em que foi publicado, pois com o avanço dos LLMs — modelos extremamente custosos de treinar e até mesmo de *fine-tunar* — ele propõe uma forma de mantê-los atualizados, oferecendo escalabilidade e reduzindo alucinações.

O artigo também explica as principais variantes de RAG utilizadas, como *Naive RAG*, *Advanced RAG* e *Modular RAG*. No *Naive RAG*, há três etapas principais: *indexing* (extração, limpeza e segmentação de documentos em *chunks*), *retrieval* (seleção dos *chunks* mais relevantes para a pergunta) e *generation* (síntese do prompt final a partir da busca e dos documentos selecionados). O *Advanced RAG* acrescenta melhorias, como o *pre-retrieval*, em que a *query* inicial é refinada, e o *post-retrieval*, que envolve *reranking* dos *chunks* e compressão do contexto. Já o *Modular RAG* permite a introdução de módulos especializados com funções distintas, como um módulo específico para *reranking* ou uma etapa de *feedback* capaz de avaliar a resposta gerada e disparar uma nova busca, se necessário.

As arquiteturas e métodos de integração são discutidos de forma detalhada, mas o artigo dedica pouca atenção a aspectos práticos como custo computacional, escalabilidade em cenários industriais e limitações das métricas automáticas de avaliação — fatores críticos para a aplicação real de sistemas RAG em larga escala.

A avaliação do RAG é abordada principalmente no contexto de tarefas de *Question Answering (QA)*, destacando pontos como a qualidade dos *scores* usados para selecionar documentos, a fidelidade e relevância das respostas, além da robustez frente a ruídos e informações incorretas. Esses aspectos revelam a complexidade de medir a efetividade do RAG, reforçando a necessidade de desenvolver métodos de avaliação mais completos e realistas, que considerem tanto a performance em laboratório quanto os desafios práticos em ambientes de produção.