Retrieval-Augmented Generation (RAG) é uma técnica que une modelos de linguagem natural com recuperação de informações.

Essa fusão permite respostas mais precisas e atualizadas, aproveitando dados dinâmicos para enriquecer o contexto.

Uma das grandes vantagens do RAG é sua capacidade de trabalhar com dados externos, como bases de conhecimento e bancos de dados.

Isso o torna muito útil em setores como saúde, direito e atendimento ao cliente.

Apesar dos benefícios, o RAG também apresenta desafios, como a necessidade de uma boa indexação vetorial e o risco de vieses nos dados.

Estratégias de chunking e pipelines bem estruturados são fundamentais para garantir resultados consistentes.

Em conclusão, o RAG representa um passo importante na evolução dos modelos de linguagem.

Ele permite sistemas mais confiáveis e explicáveis, que podem aproveitar o melhor dos dois mundos: conhecimento treinado e dados vivos.