Arquitetura RAG

1. Definição

RAG (Retrieval Augmented Generation) combina recuperação de dados com geração de linguagem.

2. Pipeline

1. Pergunta do usuário
2. Conversão em embedding
3. Busca vetorial
4. Envio do contexto ao LLM
5. Geração da resposta

3. Vantagens

- Redução de alucinações
- Respostas baseadas em fontes
- Controle do conhecimento

4. Casos de Uso

- Chat acadêmico
- Suporte técnico
- Assistentes corporativos

5. Limitações

- Dependência da qualidade dos documentos
- Latência

6. Conclusão

RAG é uma das arquiteturas mais usadas em IA aplicada atualmente.