

RAG sem Vetores para Domínios Técnicos: desenvolvimento, validação e análise de documentos agronômicos baseado em indexação hierárquica e raciocínio sobre árvore documental

Vectorless RAG for Technical Domains: development, validation and analysis of agronomic documents based on hierarchical indexing and reasoning over document tree

Eduardo Vinícius dos Santos Menezes – eduardo.santostm1202@gmail.com
Caio Saraiva Coneglian – caio.coneglian@gmail.com

INTRODUÇÃO

Os Modelos de Linguagem de Grande Escala (LLMs) geram afirmações plausíveis, porém factualmente incorretas, fenômeno designado alucinação (Ji et al., 2023). No domínio agronômico isso é risco operacional direto: erros em doses de corretivos ou em índices de saturação acarretam prejuízo econômico e dano à lavoura. A Geração Aumentada por Recuperação (RAG) mitiga o problema ao fundamentar a resposta em fontes externas recuperadas na inferência (Lewis et al., 2020); contudo, na formulação vetorial predominante, a fragmentação dos documentos em chunks rompe a hierarquia lógica do texto técnico e fragiliza a rastreabilidade das citações (Gao et al., 2023), requisito crítico onde o agrônomo responde tecnicamente pela prescrição. O RAG sem vetores substitui os embeddings por indexação estruturada e por recuperação conduzida pelo raciocínio do LLM sobre a estrutura do documento; o framework PageIndex (Zhang; Tang, 2025) materializa esse paradigma, mas foi validado apenas sobre relatórios financeiros em inglês. Este trabalho desenvolve, implementa e valida o Agronomy Assistant, assistente para análise de prescrições agronômicas baseado em RAG sem vetores, e o compara ao RAG vetorial denso e híbrido em corpus agronômico em português, sob a hipótese de que a indexação hierárquica oferece rastreabilidade superior sem perda de corretude. A pesquisa é aplicada, experimental e quantitativa.

DESENVOLVIMENTO

O corpus reúne documentos técnico-agronômicos em português: boletins técnicos da EMBRAPA, manuais de prescrição e laudos de análise de solo de uma propriedade parceira, em PDF com hierarquia explícita de seções. Comparam-se três arquiteturas: RAG vetorial denso (BGE-M3 e busca por vizinhos aproximados em Qdrant), RAG vetorial híbrido (BGE-M3 e BM25 fundidos por Reciprocal Rank Fusion) e RAG sem vetores (PageIndex com GPT-4o via Azure OpenAI). Nesta última, sem etapa de embedding, o documento é representado como árvore de nós com título, resumo e intervalo de páginas, percorrida de cima para baixo pelo raciocínio do modelo. Foi implantado em produção o Agronomy Assistant, sistema full-stack em FastAPI e React 19 com painel de citações e dashboard de métricas, que mantém as três arquiteturas ativas. A avaliação emprega ground truth sintético de 120 pares pergunta-fonte, gerados por modelo distinto do usado nas consultas para mitigar a circularidade avaliativa, em quatro categorias: factual, numérico-quantitativa, multi-hop e comparativa. Nos ensaios preliminares, com 30 perguntas das duas primeiras categorias sobre cinco documentos e corretude binária, o PageIndex citou documento, página e seção em 100% das respostas, enquanto as vetoriais citaram apenas o chunk recuperado; a corretude foi similar entre as três no subconjunto factual e a latência e os tokens do PageIndex ficaram em cerca do dobro dos do RAG denso. Registra-se como ameaça à validade a assimetria de infraestrutura entre os modelos geradores.

CONCLUSÃO

Os ensaios preliminares respondem parcialmente de forma afirmativa à pergunta central: o RAG sem vetores oferece rastreabilidade consistente até o nível de documento, página e seção, superando qualitativamente o RAG vetorial no critério mais crítico para a auditabilidade agrônômica, ao custo de maior latência e consumo de tokens. As contribuições são o sistema em produção, um ground truth sintético de 120 perguntas e um protocolo de avaliação em duas fases. Seguem-se a avaliação completa com o framework RAGAS (Es et al., 2024), a validação do ground truth por especialistas e a unificação do modelo gerador.

REFERÊNCIAS

ES, S. et al. RAGAS: automated evaluation of retrieval augmented generation. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL), 2024. Proceedings [...]. [S.l.: s.n.], 2024. p. 150-158.

GAO, Y. et al. Retrieval-augmented generation for large language models: a survey. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2312.10997>. Acesso em: 14 abril 2026.

JI, Z. et al. Survey of hallucination in natural language generation. ACM Computing Surveys, v. 55, n. 12, p. 1-38, 2023.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, v. 33, p. 9459-9474, 2020.

ZHANG, M.; TANG, Y. PageIndex: document index for vectorless, reasoning-based RAG. [S.l.]: VectifyAI, 2025. Disponível em: <https://github.com/VectifyAI/PageIndex>. Acesso em: 1 maio 2026.