APLICAÇÃO DE ABORDAGEM DE AGENTES BASEADOS EM MODELOS DE LINGUAGEM DE GRANDE ESCALA PARA MELHORAR AS RESPOSTAS NO CONTEXTO DE INFORMAÇÕES AGROPECUÁRIAS

Felipe Kenji Nishioka Jojima, Fabiano Baldo

INTRODUÇÃO

A popularização do uso de Modelos Grandes de Linguagem (*Large Language Models* - LLM) em tarefas de Processamento de Linguagem Natural (*Natural Language Processing* - NLP), decorrente da sua capacidade de geração de textos complexos, vem aumentando consideravelmente sua aplicação em situações reais (Huang et al., 2026). Entretanto, apesar do extenso treinamento desses modelos, eles ficam limitados ao conhecimento adquirido até a data em que foram treinados, chamado de conhecimento paramétrico. Portanto, eles começam a se tornar desatualizados a partir deste momento, o que pode potencializar a ocorrência de problemas de alucinações (Huang & Huang, 2026). Visando mitigar esses problemas, a arquitetura RAG (*Retrieval Augmented Generation*) (Lewis et al., 2020) propõe que sejam oferecidas bases de conhecimento externas ao LLM, chamadas de conhecimento não paramétrico, as quais fornecem informações de contexto a ele, reduzindo alucinações e necessidade de atualizá-lo constantemente (JIANG et al., 2025). Dentre as arquiteturas derivadas do RAG, destaca-se o *Agentic RAG*, que define fluxos entre agentes que executam funções distintas, o que provê uma estruturação mais robusta de verificação e validação de respostas, assim como evita a execução de processos desnecessários na geração das respostas.

Dessa maneira, o objetivo neste trabalho foi desenvolver uma arquitetura de RAG agêntico para lidar com a geração de respostas a perguntas no contexto agropecuário catarinense a partir do conhecimento não paramétrico provido pela base de dados chamada EpagriTec, construída e mantida pelo corpo técnico da Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina – EPAGRI.

DESENVOLVIMENTO

A arquitetura RAG proposta foca na otimização da recuperação e re-ranqueamento de documentos por meio de um fluxo de *fallbacks* sequenciais. O fluxo inicia expandindo a pergunta do usuário em cinco *queries* complementares, buscando um total de 30 documentos (cinco por pergunta) em um banco vetorial. A recuperação é realizada utilizando até três estratégias: a) similaridade com limiar alto (0,9), chamada de RS1; b) similaridade sem limiar, chamada de RS2; e c) MMR (*Maximal Marginal Relevance*) com fator de diversidade (0,5), chamada de RS3. Após a busca, os documentos são submetidos a uma métrica de *Context Relevance*, aprovando apenas aqueles com nota superior a 0,6 e guardando esse histórico entre os ciclos. As respostas geradas a partir desse escopo também passam por um crivo estrito, devendo atingir limiares de relevância (0,9) e de utilização de contexto (0,85). Caso os documentos recuperados ou a resposta não atinjam o limite para aprovação em um ciclo, ele avança automaticamente para a próxima estratégia de busca, retornando erro ao usuário apenas se falhar na tentativa do último ciclo (MMR).

Para a validação prática do fluxo, construiu-se uma base de dados vetorial sobre cultivos de arroz e banana via *web scraping* do site EpagriTec utilizando o *Multilingual-E5-Large* para embeddings. O experimento foi implementado utilizando *LangChain* e *LangGraph*. Os modelos quantizados foram carregados utilizando o *Ollama* em uma infraestrutura local com GPU dedicada de 24GB NVIDIA L4: *Gemma4:26b-A4B* para geração e *Qwen3.5:9B* para as funções avaliadoras. O rigor do sistema foi suportado pelos frameworks *RAGAs* e *BERTScore*, garantindo que o agente recuse informações de baixa qualidade em vez de alucinar as respostas.

RESULTADOS

A avaliação do fluxo RAG multi-estratégia foi realizada sobre um conjunto de 8 perguntas formuladas por especialistas (Q1 a Q8), submetidas a 30 execuções cada, a fim de demonstra a eficácia do mecanismo de *fallback* e a distribuição do acionamento das diferentes estratégias de busca. Conforme ilustrado na Figura 1, questões mais abrangentes como Q1 e Q2 foram integralmente resolvidas pela estratégia inicial (RS1), evidenciando que a base possui documentos com proximidade vetorial suficiente para estes temas. Em contrapartida, perguntas que exigiam a correlação de dados dispersos, como a Q3, demandaram majoritariamente da ativação da estratégia RS3. A pergunta Q5 revelou-se o caso mais complexo, resultando em respostas não atendidas na maioria das inferências por exigir do sistema a identificação de informações não descritas de forma explícita e direta nos documentos.

A análise da relevância das respostas geradas evidencia o impacto positivo das múltiplas estratégias de recuperação disponíveis, comportamento ilustrado na Figura 2. O boxplot revela uma tendência de aumento nos valores medianos da métrica de *Answer Relevancy* à medida que o sistema avança da RS1 para a RS3 no grupo de questões que ativou os fallbacks (Q3, Q5, Q6 e Q8). O teste estatístico par-a-par de *Wilcoxon*, aplicado com correção de Holm-Bonferroni a 5% de significância, confirmou que as diferenças de desempenho entre as três estratégias são estatisticamente significativas. A RS3, ao garantir maior diversidade (saltando, por exemplo, de uma média de 5,41 na RS1 para 14,59 documentos aprovados na Q3), apresentou capacidade superior para embasar respostas que dependem da integração de múltiplos textos.

A validação da fidelidade semântica em relação às respostas de referência, medida pelo BERTScore, apresentou valores consistentes de F1-Score variando entre 0,8467 (Q4, RS1) e 0,9445 (Q7, RS2). Um padrão observado em todas as questões foi a prevalência da métrica de Cobertura sobre a Precisão, indicando que as respostas geradas pelo modelo tendem a abranger todo o conteúdo semântico esperado pelo especialista, mesmo quando incluem termos adicionais não presentes na resposta de referência. Para as questões mais complexas, observou-se que o uso da RS3 não comprometeu a qualidade da resposta e, em casos como Q5 e Q8, efetivamente aprimorou o F1-Score em comparação com a busca inicial, confirmando que a arquitetura consegue manter estabilidade e coerência limitando o ruído por meio das avaliações em tempo de execução.

CONSIDERAÇÕES FINAIS

O estudo demonstra que o fluxo RAG agêntico é uma solução robusta para domínios especializados (como o agrícola), evidenciando que a simples recuperação de documentos é insuficiente para evitar respostas incorretas ou incompletas. A integração de avaliações *ground-truth-free* (sem resposta de referência) em tempo de execução e rotas de *fallback* garantiu respostas de melhor qualidade e fidelidade. A divisão do sistema em quatro fases (Expansão, Recuperação, Re-ranqueamento e Geração) conferiu a flexibilidade e adaptabilidade esperadas na geração das respostas, validada empiricamente no portal EpagriTEC. Para trabalhos futuros, sugere-se a evolução para um RAG adaptativo — com orquestração inteligente para expansão de query e ajuste dinâmico da temperatura do modelo baseados na complexidade da pergunta —, além da aplicação em *benchmarks* científicos consolidados na literatura.

Palavras-chave: Retrieval Augmented Generation; Large Language Models; Information Retrieval; Multi-strategy Retrieval;

ANEXOS

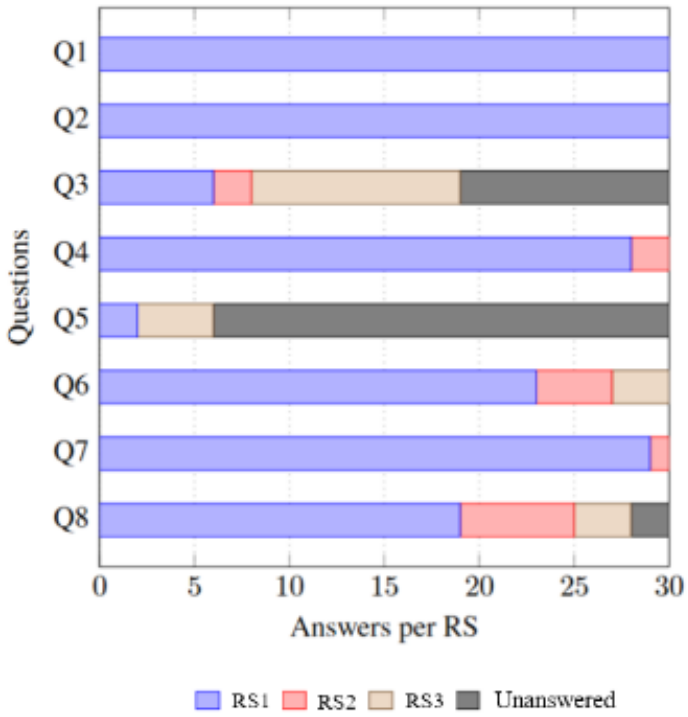


Figura 1. *Distribuição das respostas por estratégia de recuperação (Fonte: O autor).*

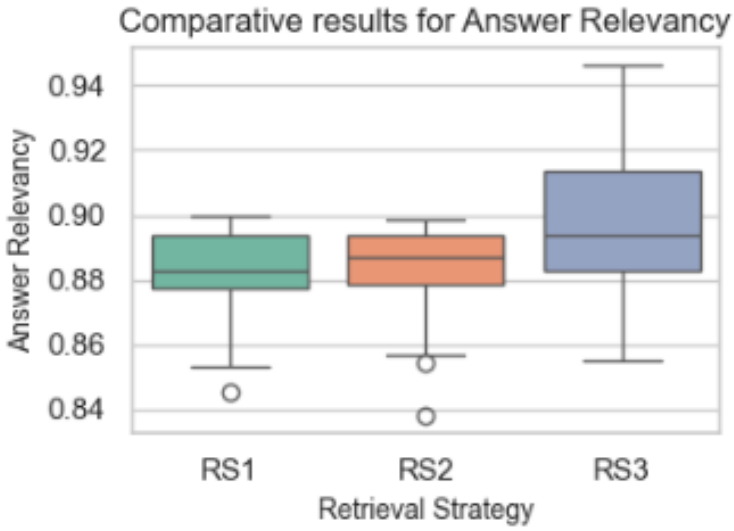


Figura 2. *Comparativo de relevância de resposta entre as estratégias de recuperação (Fonte: O autor).*

REFERÊNCIAS BIBLIOGRÁFICAS

HUANG, Yizheng; HUANG, Jimmy Xiangji. **A Survey on Retrieval-Augmented Text Generation for Large Language Models.** ACM Computing Surveys, New York, abr. 2026 . DOI: https://doi.org/10.1145/3805774. Acesso em: 01 set. 2026.

JIANG, Pengcheng; OUYANG, Siru; JIAO, Yizhu; ZHONG, Ming; TIAN, Runchu; HAN, Jiawei. Retrieval And Structuring Augmented Generation with Large Language Models. In: **Proceedings Of The 31st Acm Sigkdd Conference On Knowledge Discovery And Data Mining V.2 (Kdd '25)**, 2025, Toronto, ON, Canada. New York: ACM, 2025. p. 1-11. DOI:https://doi.org/10.1145/3711896.3736557. Acesso em: 01 set. 2026.

HUANG, Yujiao; YANG, Ling; YANG, Xu-Hua; XU, Xinli. **Retrieval-Augmented Generation for Multi-Hop Question Answering Based on Structured Planning**. ACM Transactions on Knowledge Discovery from Data, LOCAL DE PUBLICAÇÃO NÃO ENCONTRADO, v. 20, n. 3, Artigo 44 (20 páginas), 2026. https://doi.org/10.1145/3789506. Disponível em: https://dl.acm.org/doi/10.1145/3789506. Acesso em: 01 set. 2026.

LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KUTTLER, H.; LEWIS, M.; YIH, Wen-Tau; ROCKTASCHEL, Tim; RIEDEL, S.; KIELA, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *In*: **NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems**, n. 793, p. 9459 - 9474. Disponível em: https://dl.acm.org/doi/abs/10.5555/3495724.3496517. Acesso em: 01 set. 2026.