APLICAÇÃO DE TÉCNICAS DE RETRIEVAL-AUGMENTED GENERATION E ENGENHARIA DE PROMPT PARA MELHORAR AS RESPOSTAS DE MODELOS DE LINGUAGEM DE GRANDE ESCALA NO CONTEXTO DE INFORMAÇÕES AGROPECUÁRIAS

Vinícius Giroti, Fabiano Baldo

INTRODUÇÃO

A crescente adoção de tecnologias no setor agrícola tem impulsionado a demanda por acesso rápido a informações técnicas especializadas (Oliveira e Silva, 2023). Contudo, apesar dos avanços dos mecanismos de busca a partir da adoção de Processamento de Linguagem Natural (Natural Languege Processing - NLP), impulsionada pelo uso de Modelos Grandes de Linguagem (Large Language Models - LLMs), eles ainda apresentam limitações relacionadas à geração de respostas a perguntas técnicas relacionadas a domínios específicos (Huang et al., 2025). Dessa maneira, este trabalho tem como objetivo complementar o conhecimento paramétrico das LLMs por meio da adição de conhecimento não paramétrico obtido a partir da aplicação de técnicas de Retrieval Augmented Generation (RAG) (Lewis et al., 2020) para geração de respostas a perguntas relacionadas ao contexto agropecuário catarinense. Portanto, neste trabalho são agregados conteúdos agrícolas textuais e descritivos de imagens ao conhecimento da LLM, utilizando como base de dados a plataforma EpagriTec. O foco central da pesquisa é investigar como o enriquecimento multimodal da base textual a partir das descrições automatizadas de imagens contidas na plataforma EpagriTec afetam a recuperação de informações e a fidelidade das respostas de LLMs no domínio agropecuário catarinense.

DESENVOLVIMENTO

A proposta deste trabalho consistiu na construção de um fluxo RAG dividido em duas etapas: 1) Etapa de Ingestão, que incorpora a descrição textual das imagens contidas nos documentos da plataforma EpagriTec durante o processo de vetorização (embedding) e posterior armazenamento no banco de dados vetorial; 2) Etapa de Inferência, onde a recuperação de informações relevantes é realizada no banco de dados vetorial e incorporada ao contexto da LLM para a geração da resposta à pergunta do usuário. Na etapa de ingestão, empregou-se o modelo Gemini Pro 2.5 (Comanici et al., 2025) para a geração das descrições textuais das imagens contidas na plataforma, que foram integradas aos documentos originais, os quais foram divididos em pedaços (chunks), que foram vetorizados pelo modelo BAAI/bge-m3 (Chen et al., 2025) e armazenados no banco vetorial FAISS (Facebook AI Similarity Search). Na fase de inferência, aplicou-se a engenharia de prompt para concatenar a pergunta do usuário ao contexto enriquecido recuperado para a geração da resposta pelo LLM. Os LLMs de código aberto avaliados nos experimentos foram: Gemma (Gemma Team et al., 2024), Llama 3 (Grattafiori et al., 2024) e Ministral-8B (Liu et al., 2026). Todos foram configurados com temperatura de 0,2 a fim de que eles priorizassem estritamente a síntese da resposta a partir do contexto recuperado do banco de dados vetorial (conhecimento não paramétrico), limitando assim o uso do conhecimento pré-treinado (conhecimento paramétrico) na tentativa de mitigar alucinações.

RESULTADOS

A avaliação experimental evidenciou que o enriquecimento multimodal (texto + imagens) do RAG e a estruturação do prompt superaram o déficit informacional existente na geração da resposta a partir do conhecimento paramétrico dos LLMs. A injeção das descrições de imagens aumentou a taxa de acerto na recuperação de contextos relevantes (Hit Rate@5) de 80% para 100%, como mostrado pela Tabela 2. Na avaliação da qualidade do texto gerado, mensurada pela métrica de similaridade semântica BERTScore (Zhang et al., 2020), por meio dos modelos avaliadores XLM-RoBERTa (Conneau et al., 2020) e BERTimbau (Souza et al., 2020), as médias globais de todos os LLMs testados apresentaram um aumento expressivo com diferença significativa quando as descrições das imagens estavam presentes nos documentos. O modelo Ministral-8B operando com a base de dados com descrições obteve o melhor desempenho da pesquisa, alcançando um F1-Score médio de 0,643 com o avaliador BERTimbau e 0,885 com o XLM-RoBERTa, como evidenciado pela Tabela 1. A avaliação qualitativa confirmou que o contexto enriquecido reduziu as respostas desconexas (alucinadas), embora tenha revelado que descrições automatizadas de imagens mal catalogadas podem introduzir poluição semântica e prejudicar a inferência da resposta pelo LLM.

CONSIDERAÇÕES FINAIS

A aplicação de técnicas de RAG e engenharia de prompt, especialmente mediante a injeção de descrições textuais de imagens contidas no texto original, elevam a fidelidade factual dos LLMs na geração de respostas a perguntas de domínios específicos, tais como avaliado no contexto agropecuário catarinense. Portanto, é possível inferir que o enriquecimento multimodal por descrição de imagens pode ser considerado um fator determinante para que os modelos consigam realizar a geração de respostas mais precisas em domínios específicos, mitigando assim alucinações ao embasar a geração das respostas a partir de informações com contextos semanticamente mais completos e acurados.

Palavras-chave: Agricultura Digital; Geração Aumentada por Recuperação; Modelos Grandes de Linguagem; Avaliação Semântica; Engenharia de Prompt.

ANEXOS

Tabela 1. *Resultados de Bertscore para a avaliação dos modelos sem descrição e com descrição.*

ID	Embedding	gemma (w/o Desc.)	gemma (w/ Desc.)	llama3 (w/o Desc.)	llama3 (w/ Desc.)	ministral-8b (w/o Desc.)	ministral-8b (w/ Desc.)
P1	BERTIMBAU	0.370 ± 0.002 (4)	0.595 ± 0 (3)	0.324 ± 0 (5)	0.613 ± 0.019 (1)	0.324 ± 0 (5)	0.601 ± 0.015 (2)
P2	BERTIMBAU	0.636 ± 0.016 (3)	0.668 ± 0.011 (2)	0.610 ± 0.031 (4)	0.597 ± 0.022 (6)	0.605 ± 0.033 (5)	0.669 ± 0.004 (1)
P3	BERTIMBAU	0.585 ± 0 (6)	0.605 ± 0.007 (5)	0.618 ± 0.031 (3)	0.684 ± 0.019 (2)	0.616 ± 0.037 (4)	0.718 ± 0.011 (1)
P4	BERTIMBAU	0.639 ± 0.013 (6)	0.647 ± 0.007 (5)	0.653 ± 0.008 (4)	0.661 ± 0.008 (3)	0.667 ± 0.012 (2)	0.671 ± 0.015 (1)
P5	BERTIMBAU	0.546 ± 0.001 (6)	0.560 ± 0.006 (4)	0.579 ± 0.009 (3)	0.547 ± 0.001 (5)	0.592 ± 0.002 (1)	0.587 ± 0.018 (2)
P6	BERTIMBAU	0.586 ± 0.004 (5)	0.596 ± 0.033 (3)	0.554 ± 0.002 (6)	0.642 ± 0.044 (2)	0.596 ± 0.018 (3)	0.729 ± 0.047 (1)
P7	BERTIMBAU	0.649 ± 0 (4)	0.646 ± 0.003 (6)	0.649 ± 0.001 (4)	0.652 ± 0.006 (3)	0.667 ± 0.007 (1)	0.664 ± 0.009 (2)
P8	BERTIMBAU	0.542 ± 0 (5)	0.635 ± 0.015 (1)	0.603 ± 0.027 (3)	0.498 ± 0.046 (6)	0.562 ± 0.047 (4)	0.633 ± 0.018 (2)
P9	BERTIMBAU	0.647 ± 0.013 (3)	0.613 ± 0.010 (4)	0.724 ± 0.018 (1)	0.526 ± 0 (6)	0.702 ± 0.033 (2)	0.553 ± 0.036 (5)
P10	BERTIMBAU	0.610 ± 0.001 (1)	0.608 ± 0.001 (2)	0.608 ± 0.001 (2)	0.608 ± 0 (2)	0.603 ± 0.002 (6)	0.605 ± 0.002 (5)
Overall Average	BERTIMBAU	0.581 ± 0.005	0.617 ± 0.009	0.592 ± 0.013	0.603 ± 0.017	0.593 ± 0.019	0.643 ± 0.017
Average Ranking	BERTIMBAU	4.3	3.5	3.5	3.6	3.3	2.2
P1	XLM-Roberta	0.825 ± 0.005 (6)	0.867 ± 0 (2)	0.834 ± 0 (4)	0.869 ± 0.004 (1)	0.834 ± 0 (4)	0.866 ± 0.004 (3)
P2	XLM-Roberta	0.898 ± 0.006 (3)	0.906 ± 0.004 (2)	0.881 ± 0.010 (4)	0.876 ± 0.007 (6)	0.880 ± 0.012 (5)	0.911 ± 0.001 (1)
P3	XLM-Roberta	0.884 ± 0 (6)	0.886 ± 0.001 (4)	0.886 ± 0.006 (4)	0.902 ± 0.006 (2)	0.889 ± 0.005 (3)	0.913 ± 0.002 (1)
P4	XLM-Roberta	0.870 ± 0.004 (6)	0.875 ± 0.002 (1)	0.875 ± 0.002 (1)	0.875 ± 0.002 (1)	0.877 ± 0.005 (2)	0.874 ± 0.007 (5)
P5	XLM-Roberta	0.865 ± 0 (5)	0.879 ± 0.001 (2)	0.873 ± 0.002 (4)	0.864 ± 0 (6)	0.880 ± 0.001 (1)	0.879 ± 0.007 (2)
P6	XLM-Roberta	0.873 ± 0 (3)	0.872 ± 0.008 (4)	0.863 ± 0.001 (6)	0.880 ± 0.014 (2)	0.868 ± 0.004 (5)	0.901 ± 0.012 (1)
P7	XLM-Roberta	0.894 ± 0 (1)	0.893 ± 0.001 (6)	0.894 ± 0 (1)	0.894 ± 0.001 (1)	0.894 ± 0 (1)	0.895 ± 0.001 (2)
P8	XLM-Roberta	0.853 ± 0 (6)	0.881 ± 0.005 (1)	0.866 ± 0.008 (3)	0.856 ± 0.005 (5)	0.861 ± 0.009 (4)	0.881 ± 0.003 (1)
P9	XLM-Roberta	0.873 ± 0.002 (3)	0.861 ± 0.002 (4)	0.893 ± 0.005 (1)	0.845 ± 0 (6)	0.885 ± 0.011 (2)	0.850 ± 0.006 (5)
P10	XLM-Roberta	0.883 ± 0 (3)	0.884 ± 0.001 (1)	0.884 ± 0 (1)	0.884 ± 0 (1)	0.882 ± 0.001 (5)	0.883 ± 0.001 (3)
Overall Average	XLM-Roberta	0.872 ± 0.002	0.880 ± 0.002	0.875 ± 0.004	0.874 ± 0.004	0.875 ± 0.005	0.885 ± 0.004
Average Ranking	XLM-Roberta	4.2	2.7	2.9	3.1	3.2	2.4

Tabela 2. *Comparação da avaliação dos documentos recuperados entre o banco de dados com descrição das imagens e o banco de dados sem descrição das imagens.*

ID	Pipeline	Correct	Incorrect	Retrieval Path
P1	Without Description	0	5	Reranking Search
P1	With Description	5	0	Similarity Search
P2	Without Description	2	0	Reranking Search
P2	With Description	4	0	Reranking Search
P3	Without Description	5	0	Reranking Search
P3	With Description	5	0	Reranking Search
P4	Without Description	5	0	Reranking Search
P4	With Description	5	0	Reranking Search
P5	Without Description	5	0	Reranking Search
P5	With Description	5	0	Reranking Search
P6	Without Description	0	2	Reranking Search
P6	With Description	5	0	Reranking Search
P7	Without Description	5	0	Reranking Search
P7	With Description	5	0	Reranking Search
P8	Without Description	2	0	Reranking Search
P8	With Description	5	0	Reranking Search
P9	Without Description	5	0	Reranking Search
P9	With Description	5	0	Similarity Search
P10	Without Description	2	0	Reranking Search
P10	With Description	4	0	Reranking Search

REFERÊNCIAS BIBLIOGRÁFICAS

CHEN, Jianlv et al. **M3-Embedding**: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. [S. l.: s. n.], 2025. Disponível em: https://arxiv.org/abs/2402.03216. Acesso em: 31 ago. 2026.

COMANICI, Gheorghe et al. **Gemini 2.5**: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. [S. l.: s. n.], 2025. Disponível em: https://arxiv.org/abs/2507.06261. Acesso em: 31 ago. 2026.

CONNEAU, Alexis et al. **Unsupervised cross-lingual representation learning at scale**. [S. l.: s. n.], 2020. Disponível em: https://arxiv.org/abs/1911.02116. Acesso em: 31 ago. 2026.

GEMMA TEAM et al. **Gemma**: open models based on Gemini research and technology. [S. l.: s. n.], 2024. Disponível em: https://arxiv.org/abs/2403.08295. Acesso em: 31 ago. 2026.

GRATTAFIORI, Aaron et al. **The Llama 3 herd of models**. [S. l.: s. n.], 2024. Disponível em: https://arxiv.org/abs/2407.21783. Acesso em: 31 ago. 2026.

HUANG, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. **ACM Trans. Inf. Syst.**, New York, v. 43, n. 2, jan. 2025. DOI: https://doi.org/10.1145/3703155. Disponível em: https://dl.acm.org/doi/10.1145/3703155. Acesso em: 27 ago. 2026.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, Vancouver. **Proceedings** [...]. Red Hook: Curran Associates Inc., 2020.

LIU, Alexander H. et al. **Ministral 3**. [S. l.: s. n.], 2026. Disponível em: https://arxiv.org/abs/2601.08584. Acesso em: 31 ago. 2026.

OLIVEIRA, R. C. d.; SILVA, R. D. d. S. e. Artificial intelligence in agriculture: benefits, challenges, and trends. **Appl. Sci.**, [s. l.], v. 13, p. 7405, 2023. DOI: https://doi.org/10.3390/app13137405. Disponível em: https://www.mdpi.com/2076-3417/13/13/7405. Acesso em: 27 ago. 2026.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: CERRI, Ricardo; PRATI, Ronaldo C. (ed.). **Intelligent systems**. Cham: Springer International Publishing, 2020. p. 403-417.

ZHANG, Tianyi et al. **BERTScore**: evaluating text generation with BERT. [S. l.: s. n.], 2020. Disponível em: https://arxiv.org/abs/1904.09675. Acesso em: 31 ago. 2026.