

Resumo de dissertação

***Large Language Models* para recuperação da informação em acervos históricos digitalizados**

Daiane Campos Procópio

Orientador: Renato Rocha Souza (Universidade Federal de Minas Gerais)

154 p. – Dissertação (Mestrado) – Defesa 05.12.2025

Resumo: Os avanços tecnológicos que ampliaram o acesso à informação no meio digital impulsionaram a produção científica, técnica, artística e cultural. Contudo, o grande volume de informações disponíveis trouxe novos desafios, especialmente para a recuperação de conteúdos relevantes e acessíveis a pessoas com diferentes necessidades e capacidades. Entre esses desafios, destacam-se os documentos textuais digitalizados, comuns em acervos institucionais, que muitas vezes não possuem caracteres reconhecíveis por softwares de leitura, dificultando a recuperação e o uso das informações. Nesse contexto, os *Large Language Models* (LLMs) surgem como uma tecnologia promissora por sua capacidade de gerar, resumir, traduzir, interpretar e recuperar informações, mesmo em textos não estruturados. Diante desse cenário, esta pesquisa teve como objetivo investigar o potencial de utilização dos LLMs no processo de identificação e recuperação da informação em documentos textuais digitalizados de repositórios institucionais, utilizando como amostra teses do Repositório Institucional da Universidade Federal de Minas Gerais (RI-UFMG). A pesquisa é de natureza aplicada e exploratória, com abordagem quali-quantitativa, e abrange um estudo de caso. Foram analisadas quarenta teses, cinco de cada uma das oito grandes áreas do conhecimento, defendidas entre 1962 e 2009, período em que as teses do RI-UFMG foram digitalizadas. A seleção considerou os recursos disponíveis, o caráter exploratório do estudo e a diversidade estrutural dos documentos, o que possibilitou avaliar o comportamento dos modelos em contextos heterogêneos. Os três modelos analisados foram o Mistral 7B, Llama 3.2 e Qwen 2.5. Eles foram avaliados ao longo de setembro de 2025, utilizando um computador pessoal. A análise baseou-se em uma amostra composta por 40 documentos, a partir dos quais foram geradas 600 respostas (200 por modelo) para cinco perguntas padronizadas, contemplando tanto questões objetivas quanto interpretativas. Os resultados indicaram que o Mistral 7B apresentou o melhor desempenho geral, com maior número de respostas coerentes e menor incidência de alucinações, seguido pelo Llama 3.2, enquanto o Qwen 2.5 obteve o menor índice de coerência. Observou-se que a precisão das respostas diminuiu à medida que aumenta a complexidade semântica das perguntas, indicando que os LLMs são mais eficazes em tarefas de recuperação literal do que em interpretações conceituais. A análise qualitativa evidenciou, ainda, desafios computacionais na utilização dos modelos e a influência das diferenças estruturais dos documentos sobre a qualidade das respostas. Concluiu-se que os LLMs apresentam potencial para apoiar a recuperação da informação em acervos digitalizados, desde que seu uso ocorra de forma supervisionada, com infraestrutura adequada e critérios de validação das respostas. Ressalta-se que o estudo tem caráter exploratório e não busca generalizações estatísticas. Como contribuição, a pesquisa apresenta uma metodologia replicável e adaptável a outros contextos e aponta caminhos para futuros estudos voltados à integração entre inteligência artificial e gestão da informação, favorecendo o fortalecimento da ciência aberta e da democratização do conhecimento.

Palavras-chave: recuperação da informação; grande modelos de linguagem; digitalização; reconhecimento óptico de caracteres.

Large Language Models for Information Retrieval in Digitized Historical Collections

Abstract: Technological advances that have expanded access to information in the digital environment have driven scientific, technical, artistic, and cultural production. However, the large volume of available information has brought new challenges, especially regarding the retrieval of relevant content that is accessible to people with different needs and abilities. Among these challenges are digitized textual documents, common in institutional collections, which often lack characters recognizable by reading software, hindering information retrieval and use. In this context, Large Language Models (LLMs) emerge as a promising technology due to their ability to generate, summarize, translate, interpret, and retrieve information, even from unstructured texts. Given this scenario, this study aimed to investigate the potential use of LLMs in the process of identifying and retrieving information from digitized textual documents in institutional repositories, using a sample of theses from the Institutional Repository of the Federal University of Minas Gerais (RI-UFGM). The research is applied and exploratory in nature, employs a qualitative–quantitative approach, and includes a case study. Forty theses were analyzed—five from each of the eight major areas of knowledge—defended between 1962 and 2009, the period during which the RI-UFGM theses were digitized. The selection considered the available resources, the exploratory character of the study, and the structural diversity of the documents, which allowed the evaluation of model performance in heterogeneous contexts. The three models analyzed were Mistral 7B, Llama 3.2, and Qwen 2.5. They were evaluated throughout September 2025 using a personal computer. The analysis was based on a sample of 40 documents, from which 600 responses were generated (200 per model) to five standardized questions, which included both objective and interpretative items. The results indicated that Mistral 7B achieved the best overall performance, with a higher number of coherent responses and fewer hallucinations, followed by Llama 3.2, while Qwen 2.5 presented the lowest coherence rate. The findings also showed that response accuracy decreases as the semantic complexity of the questions increases, indicating that LLMs are more effective in tasks involving literal retrieval than in conceptual interpretation. The qualitative analysis further highlighted computational challenges in using the models and the influence of structural differences among documents on the quality of the responses. It was concluded that LLMs have potential to support information retrieval in digitized collections, provided that their use occurs in a supervised manner, with adequate infrastructure and criteria for validating responses. It is important to emphasize that this is an exploratory study and does not aim for statistical generalization. As a contribution, the research presents a replicable methodology adaptable to other contexts and points to future studies focused on the integration between artificial intelligence and information management, strengthening open science and the democratization of knowledge.

Keywords: information retrieval; large language models; scanning; optical character recognition.

Large Language Models para la recuperación de información en acervos históricos digitalizados

Resumen: Los avances tecnológicos que han ampliado el acceso a la información en el entorno digital han impulsado la producción científica, técnica, artística y cultural. El gran volumen de información disponible ha traído nuevos desafíos, especialmente en lo que respecta a la recuperación de contenidos relevantes que sean accesibles para personas con diferentes necesidades y capacidades. Entre estos desafíos se encuentran los documentos textuales digitalizados, comunes en colecciones institucionales, que a menudo carecen de caracteres reconocibles por los programas de lectura, lo que dificulta la recuperación y el uso de la información. Los Large Language Models (LLMs) emergen como una tecnología prometedora debido a su capacidad para generar, resumir, traducir, interpretar y recuperar información, incluso a partir de textos no estructurados. Ante este escenario, este estudio tuvo como objetivo investigar el potencial uso de los LLMs en el proceso de identificación y recuperación de información a partir de documentos textuales digitalizados en repositorios institucionales, utilizando una muestra de tesis del Repositorio Institucional de la Universidad Federal de Minas Gerais (RI-UFGM). La investigación es de carácter aplicado y exploratorio, emplea un enfoque cualitativo–cuantitativo e incluye un estudio de caso. Se analizaron cuarenta tesis, defendidas entre 1962 y 2009, período durante



el cual las tesis del RI-UFMG fueron digitalizadas. La selección consideró los recursos disponibles, el carácter exploratorio del estudio y la diversidad estructural de los documentos, lo que permitió evaluar el desempeño de los modelos en contextos heterogéneos. Los tres modelos analizados fueron Mistral 7B, Llama 3.2 y Qwen 2.5. Fueron evaluados durante todo el mes de septiembre de 2025 utilizando una computadora personal. El análisis se basó en una muestra de 40 documentos, a partir de los cuales se generaron 600 respuestas (200 por modelo) a cinco preguntas estandarizadas, que incluyeron tanto ítems objetivos como interpretativos. Los resultados indicaron que Mistral 7B obtuvo el mejor desempeño general, con un mayor número de respuestas coherentes y menos alucinaciones, seguido por Llama 3.2, mientras que Qwen 2.5 presentó la tasa más baja de coherencia. Los hallazgos también mostraron que la precisión de las respuestas disminuye a medida que aumenta la complejidad semántica de las preguntas, lo que indica que los LLMs son más eficaces en tareas que implican recuperación literal que en aquellas que requieren interpretación conceptual. El análisis cualitativo también destacó los desafíos computacionales en el uso de los modelos y la influencia de las diferencias estructurales entre los documentos en la calidad de las respuestas. Se concluye que los LLMs tienen potencial para apoyar la recuperación de información en colecciones digitalizadas, siempre que su uso se realice de manera supervisada, con infraestructura adecuada y criterios para la validación de las respuestas. Es importante destacar que se trata de un estudio exploratorio y que no pretende una generalización estadística. Como contribución, la investigación presenta una metodología replicable y adaptable a otros contextos y señala la necesidad de futuros estudios centrados en la integración entre la inteligencia artificial y la gestión de la información, fortaleciendo la ciencia abierta y la democratización del conocimiento.

Palabras-clave: recuperación de la información; grandes modelos de lenguaje; digitalización; reconocimiento óptico de caracteres.

Como citar esta dissertação: PROCÓPIO, Daiane Campos. **Large Language Models para recuperação da informação em acervos históricos digitalizados.** 2025. 154 f. Dissertação (Mestrado em Gestão e Organização do Conhecimento) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2025. Disponível em: <https://hdl.handle.net/1843/3339>. Acesso em: 15 jul. 2026.