



## **Defesa de Dissertação**

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

**DAIANE CAMPOS PROCOPIO**

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. 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, a partir de 600 respostas geradas (200 por modelo) a cinco perguntas padronizadas, que variaram entre questões objetivas e 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 diminui 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.

### **Comissão Examinadora**

Prof. Renato Rocha Souza - FGV/RJ (Orientador)

Prof. Carlos Henrique Marcondes - UFF/UFMG

Profa. Patricia Nascimento Silva - ECI/UFMG

Profa. Benildes Coura Moreira dos Santos Maculan - ECI/UFMG (Suplente)

**05 de dezembro de 2025**

**15:00h**

**<https://fgv-br.zoom.us/j/5816887275>**