VI FORPED PPGGOC UFMG

Fórum de Pesquisas Discentes do Programa de Pós-Graduação em Gestão e Organização do Conhecimento

ISSN: 2965-4068

MODALIDADE:
RESUMO EXPANDIDO

Daiane Campos Procópio
Mestranda do Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Universidade Federal de Minas Gerais, Brasil.
https://orcid.org/0000-0002-9006-191X
campos-daiane@ufmg.br

Patricia Nascimento Silva
Docente do Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Universidade Federal de Minas Gerais, Brasil.
https://orcid.org/0000-0002-2405-8536
patricians@ufmg.br

Renato Rocha Souza
Docente do Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Universidade Federal de Minas Gerais, Brasil.
https://orcid.org/0000-0002-1895-3905
rsouza@eci.ufmg.br

APLICAÇÃO DO MODELO GPT-4 PARA EXTRAÇÃO DE INFORMAÇÕES EM TESES DIGITALIZADAS

Application of the GPT-4 model for information extraction in digitized thesis

1 INTRODUÇÃO

Os avanços tecnológicos que ampliaram o acesso à informação em meio digital têm contribuído significativamente para o crescimento da produção científica, técnica, artística e cultural em diversos países. Esse fenômeno também é evidente nas instituições de ensino e pesquisa, onde a produção intelectual se expande continuamente. Porém, isso traz desafios, principalmente na recuperação da informação em documentos textuais digitalizados, visto que os caracteres desses documentos, muitas vezes, não são identificados por *softwares* de leitura de tela para pessoas com deficiência visual (Instituto Federal do Rio Grande do Sul, 2018).

A fim de identificar ferramentas que auxiliem o processo de recuperação da informação em documentos digitalizados, os *Large Language Models* (LLMs) surgem como uma solução promissora ao possibilitarem a identificação de caracteres. Desta forma, o objetivo geral deste estudo foi investigar a utilização do modelo GPT-4 no processo de identificação e recuperação da informação em documentos textuais digitalizados. Como objeto de estudo, foram utilizadas teses do Repositório Institucional da Universidade Federal de Minas Gerais (RI-UFMG), que inclui documentos digitalizados até o ano de 2010. O LLM escolhido para este estudo foi o GPT-4 por ser considerado o estado da arte em LLMs (Baktash; Dawodi, 2023). Destaca-se que este estudo

integra uma pesquisa de mestrado em andamento e justifica-se pela necessidade de aprimorar os processos de recuperação da informação em documentos textuais digitalizados.

2 FUNDAMENTAÇÃO TEÓRICA

Segundo Kallens, Kristensen-McLachlan e Christiansen (2023, p. 1, tradução nossa), LLMs são “arquiteturas sofisticadas de aprendizagem profunda treinadas em grandes quantidades de dados de linguagem natural, permitindo a realização de uma variedade de tarefas linguísticas”. LLMs podem gerar, resumir e traduzir textos, escrever códigos de computador, simular diálogos, fazer análise de sentimentos, categorizar textos em diferentes temas e tópicos, fazer correções gramaticais, entre outros. Também são capazes de realizar a busca semântica de palavras, identificando não apenas a correspondência exata dos termos, mas também o contexto, permitindo encontrar resultados relevantes, mesmo quando as palavras exatas não são usadas pelo usuário no momento da busca (Wei; Huang; Wang, 2025).

A Ciência da Informação, ao estudar os processos de representação e recuperação da informação, fornece subsídios teóricos para esses processos (Capurro; Hjørland, 2003). Eles estão diretamente relacionados aos LLMs, bem como à recuperação em ambientes digitais, sendo relevante seu estudo no âmbito da Ciência da Informação, área na qual esta pesquisa se insere.

3 METODOLOGIA

Esta pesquisa é aplicada e exploratória, abrange um estudo de caso e utiliza uma abordagem qualiquantitativa, visando identificar uma compreensão mais abrangente do problema, visto que isso não seria possível utilizando uma abordagem isolada (Creswell; Creswell, 2021; Gil, 2023).

Considerando o objeto de estudo, no total, há 1618 teses de doutorado, até o ano de 2010, digitalizadas e disponíveis no RI-UFMG. Para selecionar as publicações utilizadas nesta pesquisa, foi realizada uma busca no RI-UFMG a partir da filtragem pelo campo "Tipo de documento", escolhendo-se a opção “Tese de doutorado”. Os resultados foram organizados por data de publicação, em ordem

crescente, com exibição de 100 itens por página. Por fim, foram selecionadas as vinte primeiras teses listadas.

Para a avaliação da recuperação da informação com o modelo GPT-4, foram elaborados *prompts*, que são perguntas enviadas ao modelo em linguagem natural, com o objetivo de identificar seu desempenho na identificação de caracteres e na recuperação semântica e textual dos documentos digitalizados. Esses *prompts* foram formulados considerando diferentes tipos de informação presentes nas teses, como título, autor, resumo, objetivos, metodologia e conclusões. Ao todo, foram definidos cinco *prompts*, cada um com critérios específicos de avaliação, visando analisar a capacidade do modelo de extrair e interpretar informações relevantes.

As respostas geradas pelo modelo a partir dos *prompts* enviados foram analisadas sob duas perspectivas: quantitativa e qualitativa. Após a coleta dos dados, as respostas foram avaliadas utilizando as métricas: 1) quantitativa, com base na taxa de acertos (respostas corretas ou incorretas), e 2) qualitativa, por meio da análise da coerência das respostas, a fim de mensurar o desempenho do modelo na identificação e recuperação das informações nos documentos. Para apoiar a análise qualitativa, foi utilizada uma escala Likert adaptada, que classifica as respostas em cinco níveis de coerência, variando de “sem coerência” à “totalmente coerente”.

A análise dos resultados da escala Likert foi feita de forma descritiva, considerando a frequência observada em cada um dos pontos da escala. Foram utilizadas medidas de tendência central, como média, mediana e moda, para identificar o valor que melhor representa a distribuição das respostas e, assim, avaliar a consistência e a precisão do modelo nas tarefas propostas.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Entre as vinte teses analisadas utilizando-se os cinco *prompts*, totalizando cem perguntas, o modelo respondeu corretamente a 98, alcançando uma taxa de acerto de 98%[1]. Na análise qualitativa das respostas geradas pelo GPT customizado, utilizando-se a escala Likert, foi identificada a frequência no ponto 5

[1] O histórico dos *prompts* e respostas geradas pelo GPT customizado estão disponíveis no *link* https://chatgpt.com/share/676f1c17-4430-800d-9e56-05d3f2c5b5da.

desta escala, indicando que a maioria das respostas geradas pelo modelo foi totalmente coerente, considerando as perguntas enviadas e seu contexto.

O desempenho do GPT customizado demonstra a capacidade do modelo de compreender os *prompts* enviados em linguagem natural e extrair informações relevantes das teses digitalizadas, mesmo quando os textos não são legíveis por *softwares* de leitura tradicionais ou as informações não estão evidentes. Uma das teses foi redigida em inglês, mas o GPT customizado conseguiu identificar todas as respostas corretamente, mesmo com os *prompts* formulados em português. Isso evidencia mais um benefício significativo do uso do modelo, pois possibilita que pesquisadores realizem buscas por informações em publicações escritas em diversos idiomas.

5 CONSIDERAÇÕES FINAIS

Os LLMs apresentam um potencial significativo para auxiliar na recuperação de informações em diversas fontes, como imagens e documentos históricos. Contudo, para que essa tecnologia possa, de fato, cumprir seu papel, é essencial desenvolver pesquisas aprofundadas que investiguem as melhores práticas para sua implementação, levando em conta as diferentes necessidades e capacidades dos usuários.

A análise realizada neste estudo demonstra o alto desempenho do modelo GPT-4 na recuperação de informações em documentos textuais digitalizados, utilizando cinco *prompts* específicos para avaliar diferentes aspectos de sua performance. O modelo demonstrou habilidades para identificar e extrair metadados, sintetizar conteúdos, localizar objetivos, descrever metodologias e interpretar conclusões com precisão e coerência. Esses resultados mostram que o GPT-4 é uma ferramenta promissora para auxiliar na análise de grandes volumes de publicações científicas, otimizando processos como extração de dados e identificação de informações relevantes.

No entanto, há desafios a serem considerados, como a necessidade de conhecimento técnico, os custos financeiros (especialmente ao optar por modelos proprietários) e a infraestrutura tecnológica necessária. Apesar desses obstáculos, os potenciais benefícios já justificam a continuidade das investigações nessa área. Além disso, recomenda-se que futuras pesquisas incluam avaliações das respostas

geradas pelo modelo por especialistas das respectivas áreas de conhecimento abordadas nas teses digitalizadas.

REFERÊNCIAS

BAKTASH, J. A.; DAWODI, M. GPT-4: A Review on Advancements and Opportunities in Natural Language Processing. **ArXiv**, 2305.03195v1, 2023. Disponível em: https://arxiv.org/abs/2305.03195. Acesso em: 17 dez. 2024.

CAPURRO, R.; HJØRLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, p. 148-207, jan./abr. 2007. Disponível em: https://www.scielo.br/j/pci/a/j7936SHkZJkpHGH5ZNYQXnC/?format=pdf&lang=pt. Acesso em: 23 abr. 2025.

CRESWELL, J. W.; CRESWELL, J. D. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 5. ed. Porto Alegre: Penso, 2021.

GIL, A. C. **Como elaborar projetos de pesquisa**. 7. ed. Barueri: Atlas, 2023.

INSTITUTO FEDERAL DO RIO GRANDE DO SUL. Centro Tecnológico de Acessibilidade. **Ferramentas OCR** – entenda o que são e sua relação com a acessibilidade. Bento Gonçalves: CTA, 17 dez. 2018. Disponível em: https://cta.ifrs.edu.br/ferramentas-ocr-entenda-o-que-sao-como-funcionam-e-qual-sua-relacao-com-a-acessibilidade/. Acesso em: 21 nov. 2024.

KALLENS, P. C.; KRISTENSEN-MCLACHLAN, R. D.; CHRISTIANSEN, M. H. Large Language Models Demonstrate the Potential of Statistical Learning in Language. **Cognitive Science**, v. 47, n. 3, 2023. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1111/cogs.13256. Acesso em: 25 nov. 2024.

WEI, W. R.; HUANG, L.; WANG, J. J. **Retrieval-Augmented Generation for LLM Applications**: transforming search, recommendation, and AI assistants. Sebastopol, CA: O'Reilly, 2025.