LARGE LANGUAGE MODELS FOR INFORMATION RETRIEVAL IN DIGITALIZED DOCUMENTS: a study with the GPT-4 model

LARGE LANGUAGE MODELS PARA RECUPERAÇÃO DE INFORMAÇÃO EM DOCUMENTOS DIGITALIZADOS: um estudo com o modelo GPT-4

Daiane Campos Procópio | Patrícia Nascimento Silva |Renato Rocha Souza

https://doi.org/10.21747/21836671/pag23a17

Abstract: Technological advancements that have expanded access to digital information have driven scientific, technical, artistic, and cultural production. However, the vast amount of available information also presents challenges, particularly in retrieving relevant and accessible information for people with different needs and abilities. Digitized textual documents, common in institutional collections, amplify these challenges as they often lack machine-readable characters. This study investigated the use of the GPT-4 model for information retrieval in digitized documents from institutional repositories. The applied and exploratory research adopted a qualitative and quantitative approach to evaluate character recognition and semantic searches using a customized GPT. Twenty theses from the Federal University of Minas Gerais repository were analyzed using five prompts. The model achieved 98% precise and coherent responses, demonstrating high performance, although technical challenges still limit its large-scale application.

Keywords: Digitization; Generative artificial intelligence; Information retrieval; Large language models.

Resumo: Os avanços tecnológicos que ampliaram o acesso à informação em meio digital têm impulsionado a produção científica, técnica, artística e cultural. Contudo, o grande volume de informações disponíveis também apresenta desafios, especialmente para a recuperação de informações relevantes e acessíveis para pessoas com diferentes necessidades e capacidades. Documentos textuais digitalizados, comuns em acervos institucionais, amplificam esses desafios, pois muitas vezes não possuem os caracteres reconhecíveis por *softwares* de leitura. Este estudo investigou o uso do modelo GPT-4 na recuperação de informações em documentos digitalizados de repositórios institucionais. A pesquisa, de caráter aplicado e exploratório, adotou uma abordagem quali-quantitativa para avaliar o reconhecimento de caracteres e buscas semânticas em um GPT customizado. Foram analisadas 20 teses do repositório da Universidade Federal de Minas Gerais utilizando cinco *prompts*. O modelo alcançou 98% de respostas precisas e coerentes, demonstrando alto desempenho, embora desafios técnicos ainda limitem sua aplicação em larga escala.

Plavras-chave: Digitalização; Inteligência artificial generativa; Recuperação da informação; Modelos de linguagem de larga escala.

Introduction

Technological advancements that have expanded access to digital information have significantly contributed to the growth of scientific, technical, artistic, and cultural outputs in various countries. However, the vast volume of available information also poses challenges, particularly in identifying relevant and accessible content for individuals with diverse needs and capabilities. As Marcondes and Sayão (2002:43) point out, “finding relevant information is essential for it to be used”. Consequently, identifying tools that support the process of information retrieval has become vital in today’s informational landscape. In the scientific context, digital repositories have emerged as crucial tools.

Institutional repositories, developed since 2002 (ROSA and GOMES, 2010), serve as key instruments for storing and disseminating the intellectual output of educational and research institutions. Crow (2002:4) defines them as "digital collections capturing and preserving the intellectual output of a single or multi-university community", thereby promoting open access to knowledge.

According to Larson (2012:15), information retrieval "is concerned with the storage, organization, and searching of information collections". However, Baeza-Yates and Ribeiro Neto (2013) argue that this process alone does not guarantee that users' informational needs will be met. From the perspective of Information Science, the study of information retrieval seeks to identify the most effective means of representing and retrieving information within systems, aiming to fulfil users' needs at the time of search (FERNEDA, 2003).

The growing volume of data generated across various sectors of society (STATISTA, 2024), driven by technological progress, has presented increasing challenges for managing and retrieving information. This is also evident in educational and research institutions, where intellectual production continues to expand and feed into institutional repositories. A particular challenge arises in the case of digitized textual documents, which often hinder information retrieval due to the frequent absence of machine-readable characters, posing significant accessibility issues for users of screen-reading software, such as those with visual impairments (Instituto Federal do Rio Grande do Sul, 2018). In such instances, users must search the entire document manually, often relying on the table of contents as a guide.

To address this challenge, Large Language Models (LLMs) have emerged as a promising solution for enhancing information retrieval from digitized documents. These models facilitate access to embedded content by enabling character recognition and semantic understanding. This study is, therefore, guided by the following research question: can LLMs retrieve information from digitized documents?

Accordingly, the general objective of this study is to investigate the application of LLMs in the information retrieval process from digitized textual documents made available through institutional repositories. The study will focus on theses from the Institutional Repository of the Federal University of Minas Gerais, specifically those digitized up to the year 2010. The LLM employed will be GPT-4, recognized as a state-of-the-art LLM (BAKTASH and DAWODI, 2023).

The specific objectives of the study are: (1) to select digitized documents available in the Institutional Repository of the Federal University of Minas Gerais; (2) to customize a GPT model, based on GPT-4, for text identification and information retrieval in the selected documents; and (3) to evaluate the quality of the information retrieved using the customized GPT model.

This study is justified by the need to enhance information retrieval processes for digitized textual documents, particularly those in institutional repositories of Brazil's leading federal universities, which play a key role in preserving academic and scientific memory.

The expected contributions include not only advancing academic and scientific discourse on the use of LLMs in information retrieval, but also providing a practical study with the potential to benefit a wide range of institutional repository users, from researchers to members of the general public.

Information retrieval

In 1945, Vannevar Bush, a renowned MIT scientist, highlighted in his seminal paper *As We May Think* the challenges arising from the information explosion, which hindered access to scholarly output. He proposed technology-based solutions to address these issues. In his paper, Bush argued that the methods for sharing and reviewing research findings were outdated and ineffective, preventing such studies from reaching the researchers who could incorporate them into their own work, thus hampering scientific progress. According to Bush, for a record to be useful to science, it should be continuously disseminated, stored, and consulted, and it was a responsibility of the scientific community to ensure that this happened. He believed that the development of new instruments could help solve this problem. Among the devices he envisioned was a desk designed to function as a mechanized personal library: the Memex. With this device, users could store all their books, records, and communications, and the system would provide rapid and flexible access to the content (BUSH, 1945).

Following Bush's recommendations, government agencies and private companies began funding programs aimed at managing the dramatic growth of information, initially, in science and technology, and later, across all fields, contributing significantly to the modernization of the information industry and its core practices (SARACEVIC, 1996).

The term "information retrieval" is attributed to Calvin Mooers (FERNEDA, 2012). In his article *Zatocoding Applied to Mechanical Organization of Knowledge*, published in 1951, Mooers (1951:25) defined information retrieval as "the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him". According to the author, information retrieval involves both the intellectual aspects of describing and specifying information for search purposes, and the systems, techniques, or tools used to carry out this activity.

The pursuit of efficient knowledge representation and effective retrieval – an aspiration of researchers at the time – first manifested in the development of information retrieval studies and system effectiveness assessments. The goal was to discover the best way to represent information, both structurally and thematically, in order to optimize its retrieval (ARAÚJO, 2010).

Notably, the studies conducted by Cyril W. Cleverdon during the 1950s at Cranfield University, in the United Kingdom, represent a major contribution to the evaluation of information retrieval systems. Cleverdon, a librarian at the university's School of Aeronautics, investigated various systems for information representation and retrieval, applying metrics such as recall and precision to compare the effectiveness of different systems (ARAÚJO, 2009).

The field of information retrieval underwent significant evolution with the advent of personal computers and the internet, especially with the creation of the World Wide Web, by Tim Berners-Lee, in 1989. This transformation made information retrieval a mass phenomenon, with the rise of search tools such as Yahoo!, AltaVista, and Google. Companies began to use intranets for knowledge management, which created a demand for more robust information retrieval systems (STOCK and STOCK, 2013).

Despite the rapid evolution of information retrieval, significant challenges remain, as highlighted by James Allan *et al.* during the workshop held at the ACM SIGIR Forum in Massachusetts, in 2002:

> Information specialists and ordinary citizens alike are beginning to drown in information. The next generation of IR tools must provide dramatically better capabilities to learn, anticipate, and assist with the daily information gathering, analysis, and management needs of individuals and groups over time-spans measured in decades. These requirements require that the IR field rethink its basic assumptions and evaluation methodologies, because the approaches and experimental resources that brought the field to its current level of success will not be sufficient to reach the next level (ALLAN *et al.*, 2002:47).

Information retrieval has undergone significant transformations in recent years, driven by the adoption of new technological resources. The widespread use of mobile devices, for example, has brought both challenges and opportunities for information retrieval, such as improving accessibility and user experience on smaller screens, and enabling quick searches for a wide range of digital content (SOUZA and RODAS, 2020). Additionally, the semantic web provides a foundation for the development of more intelligent applications capable of better understanding data context, thereby enhancing information retrieval (LUZ, CONEGLIAN and SEGUNDO, 2019). Advances in generative artificial intelligence also hold great promise and may elevate the field to a new level, possibly fulfilling the vision once foreseen by Vannevar Bush.

Artificial Intelligence and Large Language Models

According to Kallens, Kristensen-McLachlan, and Christiansen (2023:1), LLMs are "sophisticated deep learning architectures trained on vast amounts of natural language data, enabling them to perform an impressive range of linguistic tasks". LLMs can generate, summarize, and translate texts, write computer code, simulate dialogues, perform sentiment analysis, classify texts by theme or topic, correct grammar, among other functions. They are also capable of conducting semantic searches by identifying not only exact term matches, but also the context, allowing users to retrieve relevant results even when the exact words are not used during the search (WEI, HUANG and WANG, 2025).

The term "artificial intelligence" was first coined in 1955, by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, scientists at Dartmouth College, when they proposed a research project to investigate "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (MCCARTHY *et al.*, 1955:1).

An LLM is a form of artificial intelligence (AI) based on a deep learning architecture and trained to interpret commands and generate responses through natural language processing (NLP). According to Topol (2024:76), deep learning is "[...] the subset of machine learning composed of algorithms that allow software to train itself to perform tasks by processing data networks across multiple layers". While the technology dates back to the 1940s, it has appeared under various names: cybernetics, during the 1940s-1960s,

connectionism, in the 1980s and 1990s, and finally, since 2006, deep learning (GOODFELLOW, BENGIO and COURVILLE, 2016).

NLP encompasses AI, machine learning, and linguistics. Its origins trace back to 1950, with the publication of Computing Machinery and Intelligence, by Alan Turing, a renowned researcher and pioneer in computer science. Turing also created the Turing Test, a method for evaluating a machine's ability to exhibit intelligent behavior indistinguishable from that of a human (MARTINS *et al.*, 2020). Vajjala *et al.* (2020) explain that NLP focuses on the development of systems capable of processing and understanding human language.

Although AI and NLP have been studied for over 50 years, and deep learning has advanced significantly in recent decades, several factors have contributed to the "ideal scenario" for the emergence and widespread adoption of LLMs. These include the explosion of data, the decreasing cost and increasing capacity of computational processing, advancements in deep learning, and the diversification of LLM applications (PATIL and GUDIVADA, 2024). These factors have also driven major technology companies — known as big techs — to invest heavily in financial resources, infrastructure, and specialized talent to develop high--performance LLMs.

In November 2022, OpenAI, a company focused on AI research and deployment (OPENAI, 2024a), launched ChatGPT, a chatbot capable of processing natural language commands and delivering responses in natural language with remarkable coherence and grammatical accuracy across a wide range of topics (SHAHRIAR and HAYAWI, 2023). OpenAI played a key role in popularizing LLMs by offering a free interface that allowed users to experiment with the conversational AI assistant. This strategy enabled people to become familiar with the various benefits of using this technology. As a result, OpenAI's LLM was rapidly integrated into everyday life and organizational operations, quickly being positioned as a central component in the future of work (ALAMMAR and GROOTENDORST, 2024; AMARATUNGA, 2023).

LLMs stand out as powerful tools for optimizing the information retrieval process. It is, therefore, essential to explore their applications in greater depth, as well as the societal benefits they may offer, promoting advances in research and contributing to the accessibility and the democratization of information access.

Methodology

According to Corrêa (2008:18), methodology refers to "the set of forms and methods used to carry out a given research project, the study of the techniques and procedures employed to achieve a specific goal [...]". In this sense, research methodology can be understood as the foundation that guides the path the researcher follows to achieve the objectives of a study. It is also essential for verifying the consistency and validity of the investigative process employed.

Gil (2023) classifies research based on four criteria: by the field of knowledge, according to the purpose, the general objectives, and the methods used. Based on Gil's classification, this study falls within the field of Applied Social Sciences, specifically Information Science. It is categorized as applied research, since it aims to acquire knowledge for use in a specific context, exploratory, as it seeks to broaden the understanding of the problem, clarify it, or

enable the formulation of hypotheses, and it adopts a case study approach, which allows for a contextualized analysis of the phenomenon under investigation. Furthermore, according to Creswell and Creswell (2021), this research employs a mixed-methods or quali-quantitative approach, combining both quantitative and qualitative data collection in order to achieve a broader understanding of the problem, which would not be possible through a single-method approach.

Universe and sample of the case study

The selection of the Institutional Repository of the Federal University of Minas Gerais for this study was based on its academic relevance at both national and international levels (UNIVERSIDADE..., 2024a). In addition to housing a vast collection of works from various fields of knowledge, the repository plays a crucial role in the dissemination of scientific research, promoting democratic access to knowledge and enhancing the visibility of the university's academic output.

Among the various types of publications available in the repository, doctoral theses were chosen for analysis because they represent the state of the art in scientific knowledge at the time of their publication, highlighting their historical value to the development of science. A cutoff was applied to select theses published up to the year 2010, when most of the theses and dissertations available in the Institutional Repository of the Federal University of Minas Gerais were digitized. Publications not available in the repository are part of the physical library collections of the Federal University of Minas Gerais. From 2010 to 2019, graduate works were submitted on CD-ROM and later migrated to the repository. From 2020 onward, theses and dissertations have been submitted in digital format (UNIVERSIDADE..., 2024b).

The LLM chosen for character recognition and information retrieval was GPT-4, developed by OpenAI, as it represents the state of the art in LLMs. The environment selected for developing this study was the paid version of ChatGPT, specifically the feature for creating and customizing a Generative Pre-trained Transformer (GPT). This platform was chosen due to its simple interface and the necessary features for customizing a GPT and conducting this study.

The GPT model was customized with the following specifications:

a) Name: Information Retrieval in Scanned Documents;

b) Description: retrieval of information through semantic and keyword-based searches;

c) Instructions: this model specializes in retrieving information from digitized theses by recognizing text characters. It must be capable of performing both semantic and keyword searches and providing accurate responses based on the content extracted from the digitized documents;

d) Capabilities: web browsing, whiteboard, DALL·E image generation, code interpreter, and data analysis.

The customized GPT has a technical limitation of uploading a maximum of twenty files over its entire lifespan. Therefore, to retrieve the publications, a search was conducted in the

repository following these steps: 1) in the search field, the option “Document Type” was selected, which displays a list of document types included in the repository, along with the number of items in parentheses; 2) the type “Doctoral Thesis” was selected; 3) search results were organized by “Document Date,” which refers to the year of publication, and ordered in ascending order (from the oldest to the most recent), with 100 results displayed per page. In total, 1,618 doctoral theses were available up to the year 2010. The first twenty doctoral theses retrieved from the Institutional Repository of the Federal University of Minas Gerais were selected for this study. Table 1 lists the twenty selected theses, including their identification numbers (ID) and year of publication.

Table 1 – List of the first selection of theses

ID	Title	Year
1	O Jogo de oposições na poesia de Cláudio Manuel da Costa	1973
2	Estrutura eletrônica e absorção óptica de impurezas de prata e cobre em cloreto de potássio	1976
3	Termodinâmica e dinâmica dos modelos de ising e ising num campo transverso: aplicação a sistemas ferroéletricos hidrogenados de baixa dimensionalidade	1982
4	Aplicação do efeito Mossbauer ao estudo de propriedades físicas do sistema ZrxTi1-xFe2	1985
5	Estudo das propriedades magnéticas e estruturais dos sistemas de ligas Fe-Al, Fe-Mn e Fe-Mn-Al	1986
6	Ler e escrever: variações sobre o mesmo tema	1990
7	A Traição de Penélope: uma leitura da escrita feminina da memória	1990
8	Aquisição de dados em experiências de RPE usando minicomputador de tempo real	1990
9	O Percurso dos sentidos	1991
10	A Construção de mundos na literatura não-realista	1991
11	Imunopatologia da Doença de Chagas crônica: análise da expressão idiotípica em anticorpos anti-epimastigota e estudo imunohistoquímico da lesão cardíaca	1993
12	Poeta e poesia inconfidentes: um estudo de arqueologia poética	1993
13	Transição da fecundidade e relações de gênero no Brasil	1994
14	Sabor e som: Sri Aurobindo, tradutor indiano (a busca de um centro em Auroville e Savitri)	1994
15	Linha, choque e mônada: tempo e espaço na obra tardia de Walter Benjamin	1994
16	O Infantil na literatura: uma questão de estilo	1995
17	Estudos com a calicreína urinária humana: A - Um novo método para purificação da enzima em larga escalaB - Caracterização cinética com substratos sintéticos dos tipos amida e éster, derivados da arginina N-substituída e com os inibidores aprotinina e benzamidina	1995
18	Nonequilibrium phase transitions in interacting particle systems	1996
19	A Institucionalização da pesquisa educacional no Brasil: estudo bibliométrico dos artigos publicados na Revista Brasileira de Estudos Pedagógicos 1944-74	1996
20	O SN NU objeto em português: um caso de incorporação semântica e sintática	1996

Source: compiled by the authors, 2025.

To evaluate information retrieval using the GPT-4 model, a set of prompts was developed, which refer to the natural language questions submitted by the user to the model, with the aim of assessing its performance in character recognition and in both semantic and textual information retrieval within the documents.

The responses generated by the model from the submitted prompts were analyzed by the authors from two perspectives: quantitative and qualitative. Table 2 provides a detailed overview of each prompt's identification number, the type of information targeted for extraction, the specific question posed, and the evaluation criteria applied to each prompt.

Table 2 – Prompts developed for evaluating the GPT-4 model

N.	Type of Information	Question	Evaluation Criterion
1	Title and author identification	What is the title of the thesis and the name of the author?	Metadata extraction: assesses the model's ability to identify and extract key structural metadata.
2	Content summary	Summarize the main content of the thesis in up to three sentences.	Semantic comprehension: measures the model's ability to interpret the text and provide a cohesive summary.
3	Objectives identification	What is the main objective of this thesis?	Extraction of specific information: evaluates the model's ability to locate and understand statements of purpose.
4	Methodology used	Briefly describe the methodology applied in the thesis.	Contextual retrieval: tests the model's performance in identifying and summarizing the methodology described.
5	Conclusions and results	What are the main conclusions or results presented in the thesis?	Interpretation of results: examines the model's ability to locate and interpret final sections, highlighting the study's conclusions.

Source: Compiled by the authors, 2025.

After data collection, the responses were evaluated using two metrics: 1) quantitative, by calculating the percentage of correct responses, and 2) qualitative, through the authors' analysis of the coherence of the answers, in order to assess the model's performance in identifying and retrieving information from the documents. To support the qualitative analysis, a Likert scale was applied, widely recognized for its ease of use in measuring perceptions (FEIJÕ, VICENTE and PETRI, 2020). According to Mattar and Ramos (2021), the Likert scale was developed by Rensis Likert in 1932, when the sociologist proposed a classification to measure attitudes toward social issues. Table 3 outlines the construction of the Likert scale adapted for this study, used to register the authors' perception of the coherence of the customized GPT-4 model's responses.

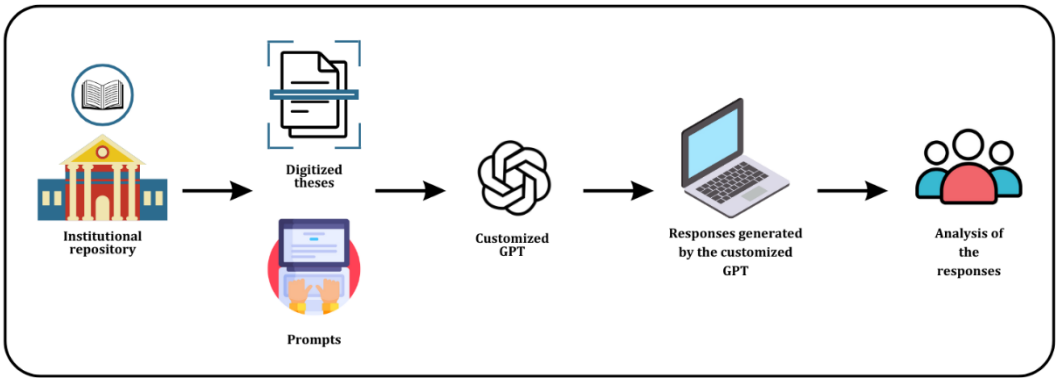
Table 3 - Likert scale adapted for this study

How coherent is the generated response in relation to the context and the question posed?	
Scale	**Description**
1	Not coherent
2	Slightly coherent
3	Moderately coherent
4	Very coherent
5	Fully coherent

Source: compiled by the authors, 2025.

Fig. 1 presents the steps outlined in the methodology adopted for the development of this study.

Fig. 1 – Steps outlined in the methodology adopted for this study



Source: compiled by the authors, 2025.

The table 4 below presents the framework used to record the responses generated by the customized GPT model. A response was classified as correct when it was free from errors and ambiguities (MICHAELIS, 2025a). Coherence was attributed to responses that demonstrated logic and relevance within the context in which they were provided (MICHAELIS, 2025b). The Appendix contains the table with the authors' detailed analysis of the responses generated by the customized GPT.

Table 4 - Record of the analysis of responses generated by the customized GPT

IDENTIFICATION		QUANTITATIVE ANALYSIS	QUALITATIVE ANALYSIS	
Thesis ID	Prompt No.	Did the model generate a correct response?	Did the model generate a coherent response?	Likert Scale assigned

Source: compiled by the authors, 2025.

The analysis of the Likert scale results was conducted descriptively, taking into account the frequency of each rating point on the adapted scale. Measures of central tendency – such as mean, median, and mode – were used to identify the value that best represented the data distribution.

Results and discussion

Customization and use of the GPT-4 model

Customizing GPT-4 requires advanced technical knowledge that goes beyond what is typically used by non-developers in their daily routines. Consequently, it was necessary to study the model and its usage in detail. The model allows only twenty files to be uploaded into its knowledge base for its entire lifecycle – ten files per upload session. When an

attempt was made to upload more documents, GPT-4 returned an error message, but without specifying the nature of the upload error. It was, therefore, necessary to consult the GPT-4 documentation to identify the issue. However, software documentation is typically written in technical language, making it difficult to understand for users without prior experience in such materials.

The documents were uploaded into the knowledge base during the GPT customization process and attached individually in the input field along with the prompts developed for model evaluation. After uploading a document, submitting the prompts, and generating the corresponding responses, a new document was uploaded for the next round.

Theses with IDs 2, 3, 4, 5, 7, 8, and 13 presented upload errors and could not be included in this study. According to GPT-4's official documentation (OPENAI, 2024b), any of the following conditions may trigger an upload error:

1. exceeding 512 MB per file;
2. reaching the limit of 2 million tokens for text files;
3. exceeding 20 MB per image file;
4. reaching the total upload limit of 10 GB per user.

Additionally, the model is not compatible with PDF files that do not contain machine-readable text (i. e., image-only scans). Based on these criteria, the likely cause of the upload errors was related to poor document scanning quality, which prevented the model from recognizing text characters, since none of the other thresholds were exceeded. As a result, seven additional theses available in the Institutional Repository of the Federal University of Minas Gerais were selected, following the same methodological criteria previously described. Table 2 details the newly selected theses:

Table 5 – List of newly selected theses

ID	Title	Year
21	Percursos críticos e tradutórios da nação: Argentina e Brasil	1996
22	Os Monstros e a questão racial na literatura pós-colonial brasileira	1997
23	Participação de antígenos do Schistosoma mansoni na indução de imunidade protetora e na modulação da reação granulomatosa	1997
24	Pontos quânticos de InAs em GaAs crescidos usando Te como surfactante	1997
25	Epidemiologia e controle da infecção por Haemophilusparasuis	1997
26	Comportamento, parâmetros fisiológicos e reprodutivos de fêmeas da raça Holandesa confinadas em freestall, durante o verão e o inverno	1997
27	O Labirinto e a bússola	1997

Source: compiled by the authors, 2025.

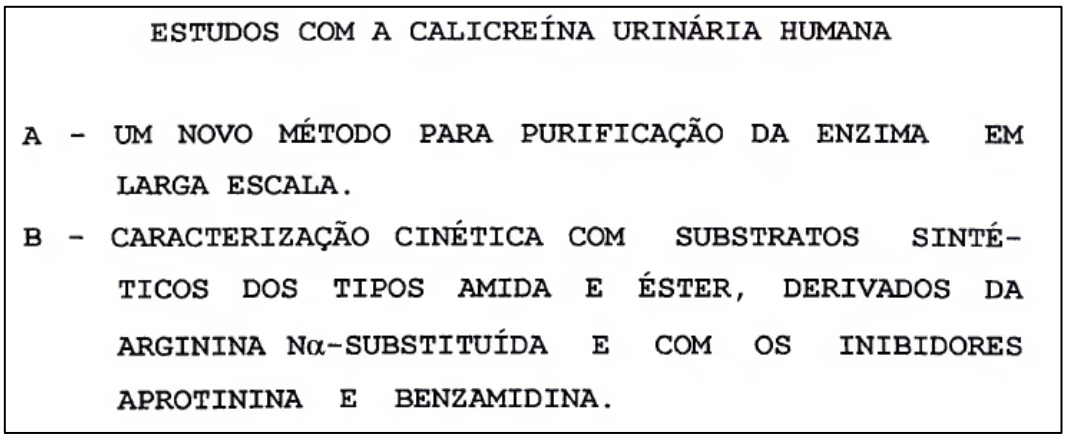
The possibility of creating ans customizing a GPT using OpenAI's GPT-4 as the base model presents a valuable opportunity for exploring diverse applications, from customer service to academic research. However, this process requires technical knowledge from the user, which can hinder the widespread adoption of this technology by individuals and organizations.

Analysis of the responses

Among the twenty theses analyzed using the five prompts described in the methodology, totaling 100 questions, the model answered 98 correctly, achieving a 98% accuracy rate[1].

The only study that yielded incoherent responses was identified with ID 17. For question number 1, "What is the title of the thesis and the name of the author?", which assessed the customized GPT's ability to extract metadata, an incomplete response was generated, as illustrated in Figures 2 and 3. This response was, therefore, marked as incorrect by the authors and assigned a rating of 1 on the Likert Scale, "Not coherent."

Fig. 2 – Title of the thesis ID 17



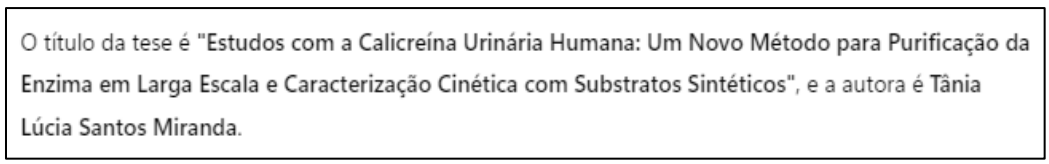
ESTUDOS COM A CALICREÍNA URINÁRIA HUMANA

A - UM NOVO MÉTODO PARA PURIFICAÇÃO DA ENZIMA EM LARGA ESCALA.

B - CARACTERIZAÇÃO CINÉTICA COM SUBSTRATOS SINTÉTICOS DOS TIPOS AMIDA E ÉSTER, DERIVADOS DA ARGININA Nα-SUBSTITUÍDA E COM OS INIBIDORES APROTININA E BENZAMIDINA.

Source: Miranda, 1995.

Fig. 3 – Response generated by the model for question 1 about thesis ID 17



O título da tese é "Estudos com a Calicreína Urinária Humana: Um Novo Método para Purificação da Enzima em Larga Escala e Caracterização Cinética com Substratos Sintéticos", e a autora é Tânia Lúcia Santos Miranda.

Source: ChatGPT, 2024.

[1] The history of the prompts and responses generated by the customized GPT is available at the following link: https://chatgpt.com/share/676f1c17-4430-800d-9e56-05d3f2c5b5da.

For question number 3, "What is the main objective of this thesis?", the customized GPT cited the specific objectives listed in the document. However, since the thesis also clearly presented a general objective, the authors deemed the response inaccurate and again assigned a rating of 1 on the Likert Scale.

Nonetheless, the customized GPT successfully identified the structural information requested in the vast majority of the theses used in this study. According to Marcondes (2005), the use of metadata on the web allows for the structured description of documents and information, assisting humans in assessing relevance and enabling software and intelligent agents to retrieve, evaluate, and process documents more effectively. Given the vast amount of data available in digital environments, well-structured metadata is essential, since simply digitizing materials does not guarantee their accessibility or usability (BACA, 2016). Furthermore, the model's strong performance in metadata identification supports the extraction of specific information, facilitating data analysis by researchers. According to Álvarez (2007), information extraction seeks to automatically identify and capture relevant data from natural language documents and structure it for output in formats such as databases or organized text.

In the qualitative analysis of the responses generated by the customized GPT, using the Likert Scale, most were rated at the highest point (5), indicating that the majority of responses were fully coherent, given the questions and their context. Table 6 presents the frequency of each Likert Scale point assigned by the authors.

Table 6 – Qualitative evaluation using the Likert scale for responses generated by the customized GPT

How coherent is the generated response in relation to the context and the question posed?	
Scale	**Number of Responses**
1	2
2	0
3	0
4	0
5	98

Source: compiled by the authors, 2025.

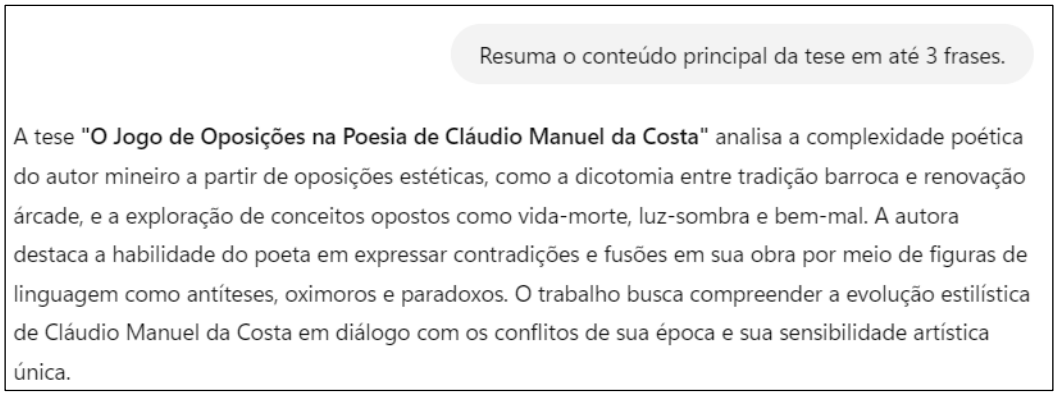
The median and mode values were both 5 – the median being the central value in the data set and the mode representing the most frequently observed value, as indicated in Table 6.

The customized GPT's performance demonstrates its ability to understand prompts written in natural language and extract relevant information from digitized theses, even when the text is not machine-readable by traditional screen readers or when the information is not explicitly stated. For example, thesis ID 27 presented its objectives under the heading "Proposed Work" rather than using the standard terms "general objective" and "specific objectives", yet the customized GPT correctly identified the content. Additionally, thesis ID 18 was written in English, but the GPT, prompted in Portuguese, accurately responded to all questions. This highlights another key benefit of the model: enabling multilingual search and retrieval across publications.

Because natural language is inherently ambiguous — a single phrase can have multiple meanings — it is essential to consider context to reduce this ambiguity (MACULAN, 2020). Recent advances in NLP have supported research in tasks such as text classification, semantic analysis, and information extraction, which are directly related to the field of Information Science. In this context, Computer Science provides innovative tools that allow Information Science to automate and enhance processes such as the creation, indexing, storage, and dissemination of information (FALCÃO, LOPES and SOUZA, 2022).

The summaries generated by the customized GPT were limited to three sentences, as requested in the corresponding prompt. The model's ability to synthesize content from scientific sources represents a valuable resource for researchers needing to analyze large volumes of academic publications, streamlining the identification of relevant materials. According to Altounian and Gomes (2016:31), semantic retrieval allows for "an understanding of concepts in their context and purpose" and is grounded in linguistics and Information Science to increase the quantity of retrieved information and improve contextual analysis through the use of natural language structures and tools to represent semantic and conceptual relationships.

Fig. 4 – Example of prompt no. 2 submitted for thesis ID 01 and the model's generated response



Source: compiled by the authors, 2025.

Two primary challenges emerged during the development of this study. The first was the lack of standardization across theses, which complicated the authors' task of locating information and analyzing the model's responses. The second challenge was the wide variety of research topics, which required the authors to invest additional effort in understanding the content and validating the coherence of the model's responses.

Despite its strong performance, technical limitations that hinder large-scale implementation persist, particularly with low-quality scanned documents. Thus, this study reinforces the importance of continuing to explore and refine the use of LLMs like GPT-4 to meet information retrieval demands in academic and scientific contexts.

Conclusion

LLMs show significant potential to support information retrieval across a wide range of sources, including images and historical documents. However, for this technology to fulfil its role effectively, in-depth research is needed to explore best practices for implementation, taking into account users' diverse needs and capabilities. This study aimed to investigate the use of LLMs in the process of retrieving information from digitized textual documents available in institutional repositories. Digitized theses from the Institutional Repository of the Federal University of Minas Gerais were used as the study's object of analysis.

The results demonstrate GPT-4's high performance in retrieving information from digitized textual documents, evaluated through five specific prompts targeting different aspects of the model's capabilities. The model successfully identified and extracted metadata, synthesized content, located research objectives, summarized methodologies, and interpreted conclusions with precision and coherence. These findings suggest that GPT-4 is a promising tool for assisting in the analysis of large volumes of academic publications, enhancing processes such as data extraction and the identification of relevant information.

As a potential solution for difficulties in recognizing characters in low-quality scans, it is proposed that optical character recognition (OCR) tools be used to preprocess scanned documents, allowing GPT-4 to subsequently retrieve their content more effectively.

The integration of LLMs such as GPT-4 into institutional repositories could offer a valuable resource for both researchers and the broader public. This integration would enable faster, more accurate searches across large volumes of academic content, facilitating knowledge access and optimizing the time needed to locate and analyze relevant information. However, several challenges must be addressed, such as technical knowledge requirements, financial costs (especially when using proprietary models), and infrastructure demands. Despite these limitations, the potential benefits justify continued research in this area, and further investigations are encouraged.

Acknowledgments

This research was carried out with the support of the Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências bibliográficas

ALAMMAR, Jay; GROOTENDORST, Maarten
2024 *Hands-On Large Language Models: language understanding and generation*. Sebastopol, CA: O'Reilly, 2024.

ALLAN, James [et al.]
2003 Challenges in information retrieval and language modeling: report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *ACM SIGIR Forum*. [Online]. 37:1 (2003) 31-47. [Retrieved 14 Aug. 2024]. Available at: https://dl.acm.org/doi/10.1145/945546.945549.

ALTOUNIAN, Márcia Martins de Araújo; GOMES, Beatriz Pinheiro de Melo
2016 A Recuperação semântica da informação no contexto do controle externo. *Revista do TCU*. [Online]. 137 (2016) 31-41. [Retrieved 22 Dec. 2024]. Available at: https://revista.tcu.gov.br/ojs/index.php/RTCU/article/view/1376/1522.

ÁLVAREZ, Alberto Cáceres
2007 *Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem*. [Online]. São Carlos, 2007. [Retrieved 21 Dec. 2024]. Available at: https://teses.usp.br/teses/disponiveis/55/55134/tde-21062007-144352/pt-br.php.
Master dissertation in Computer Sciences and Computational Mathematics - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

AMARATUNGA, Thimira
2023 *Understanding Large Language Models: learning their underlying concepts and technologies*. Nugegoda: Apress, 2023.

ARAÚJO, Carlos Alberto Ávila
2010 O Conceito de informação na Ciência da Informação. *Informação & Sociedade: Estudos*. [Online]. 20:3 (2010) 95-105. [Retrieved 31 Jul. 2024]. Available at: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/6951/4808.

ARAÚJO, Carlos Alberto Ávila
2009 Correntes teóricas da ciência da informação. *Ciência da Informação*. [Online]. 38:3 (2009) 192-204. [Retrieved 9 Jan. 2024]. Available at: https://revista.ibict.br/ciinf/article/view/1240.

BACA, Murtha
2016 Introduction. In *Introduction to Metadata*. Ed. Murtha Baca. 3ª ed. Los Angeles: Getty Publications, 2016.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier
2013 *Recuperação da informação: conceitos e tecnologia das máquinas de busca*. 2nd ed. Porto Alegre: Bookman, 2013.

BAKTASH, Jawid Ahmad; DAWODI, Mursal
2023 Gpt-4: A Review on advancements and opportunities in Natural Language Processing. *arXiv*. [Online]. 2305.03195v1, 2023. [Retrieved 17 Dec. 2024]. Available at: https://arxiv.org/abs/2305.03195.

BUSH, Vnnevar
1945 As we may think. *Atlantic Monthy*. [Online]. 176:1 (1945) 101-108. [Retrieved 18 Jan. 2024]. Available at: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

CHATGPT
2024 Recuperação de informações em digitalizações. *Prompts e respostas*. [Online]. 2024. [Retrieved 12 Jan. 2025]. Available at: https://chatgpt.com/share/676f1c17-4430-800d-9e56-05d3f2c5b5da.

CORRÊA, Luiz Nilton
2008 *Metodologia científica: para trabalhos acadêmicos e artigos científicos*. Florianópolis: Ed. do autor, 2008.

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

CROW, Raym
2002 *The Case for institutional repositories: A SPARC position paper*. [Online]. Washington, DC: The Scholarly Publishing and Academic Resources Coalition, 2002. [Retrieved 16 Aug. 2024]. Available at: https://ils.unc.edu/courses/2014_fall/inls690_109/Readings/Crow2002-CaseforInstitutionalRepositoriesSPARCPaper.pdf.

FALCÃO, Luander Cipriano de Jesus; LOPES, Brenner; SOUZA, Renato Rocha
2022 Absorção das tarefas de processamento de Linguagem Natural (NLP) pela Ciência da Informação (CI): uma revisão da literatura para tangibilização do uso de NLP pela CI. *Em Questão*. [Online]. 28:1 (2022) 13-34. [Retrieved 12 Jan. 2025]. Available at: https://doi.org/10.19132/1808-5245281.13-34.

FEIJÓ, Amanda Monteiro; VICENTE, Ernesto Fernando Rodrigues; PETRI, Sérgio Murilo
2020 O Uso das escalas Likert nas pesquisas de contabilidade. *Revista Gestão Organizacional*. [Online]. 13:1 (2020) 27-41. [Retrieved 7 Jan. 2025]. Available at: https://bell.unochapeco.edu.br/revistas/index.php/rgo/article/view/5112.

FERNEDA, Edberto
2012 *Introdução aos modelos computacionais de recuperação de informação*. Rio de Janeiro: Ciência Moderna, 2012.

FERNEDA, Edberto
2003 *Recuperação de informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. [Online] São Paulo, 2003. [Retrieved 9 Jan. 2024]. Available at: https://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/pt-br.php.
PhD thesis in Information Science and Documentation - Escola de Comunicação e Artes, Universidade de São Paulo.

GIL, Antônio Carlos
2023 *Como elaborar projetos de pesquisa*. 7ª ed. Barueri: Atlas, 2023.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron
2016 *Deep learning*. [Online]. Cambridge: MIT Press, 2016. [Retrieved 24 Jan. 2024]. Available at: https://www.deeplearningbook.org/.

INSTITUTO FEDERAL DO RIO GRANDE DO SUL. Centro Tecnológico de Acessibilidade
2018 *Ferramentas OCR: entenda o que são e sua relação com a acessibilidade*. [Online]. Bento Gonçalves: CTA, 2018. [Retrieved 21 Nov. 2024]. Available at: https://cta.ifrs.edu.br/ferramentas-ocr-entenda-o-que-sao-como-funcionam-e-qual-sua-relacao-com-a-acessibilidade/.

KALLENS, Pablo Contreras; KRISTENSEN-MCLACHLAN, Ross Deans; CHRISTIANSEN, Morten H.
2023 Large Language Models demonstrate the potential of statistical learning in language. *Cognitive Science*. [Online]. 47:3 (2023). [Retrieved 23 Aug. 2024]. Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1111/cogs.13256.

LARSON, Ray R.
2012 Information Retrieval Systems. In *Understanding Information Retrieval Systems: management, types, and standards*. Ed. Marcia J. Bates. Boca Raton: CRC Press, 2012.

LUZ, Larissa Pavarini da; CONEGLIAN, Caio Saraiva; SEGUNDO, José Eduardo Santarem
2019 Tecnologias da web semântica para a recuperação da informação no Wikidata. *Revista Digital de Biblioteconomia e Ciência da Informação*. [Online]. 17:e019003 (2019) 1-20. [Retrieved 9 Jan. 2025]. Available at: https://doi.org/10.20396/rdbci.v17i0.8651791.

MACULAN, Benildes Coura Moreira dos Santos
2020 Ambiguidade e o contexto na representação de informações em domínios de especialidade. *Perspectivas em Ciência da Informação*. [Online]. 25:número especial (2020) 98-124. [Retrieved 12 Jan. 2025]. Available at: https://periodicos.ufmg.br/index.php/pci/article/view/22284.

MARCONDES, Carlos Henrique
2005 Metadados: descrição e recuperação de informações na web. In *Bibliotecas digitais: saberes e práticas*. Org. Carlos Henrique Marcondes *et al*. Salvador: UFBA; Brasília: IBICT, 2005, p. 97-113.

MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando
2002 Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. *Ciência da Informação*. [Online]. 31:3 (2002) 42-54. [Retrieved 16 Aug. 2024]. Available at: https://www.scielo.br/j/ci/a/NKhjHgVf63bYGmkHJWQkWhB/?format=pdf&lang=pt.

MARTINS, Júlio Serafim [et al.]
2020 *Processamento de linguagem natural*. Porto Alegre: SAGAH, 2020.

MATTAR, João; RAMOS, Daniela Karine
2021 *Metodologia da pesquisa em educação: abordagens qualitativas, quantitativas e mistas*. São Paulo: Almedina Brasil, 2021.

MCCARTHY, John [et al.]
1955 A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*. [Online]. 27:4 (1955) 12. [Retrieved 29 Mar. 2024]. Available at: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.

MICHAELIS
2025a *Dicionário Brasileiro da Língua Portuguesa*. [Online]. São Paulo: Melhoramentos, 2025. [Retrieved 15 Jan. 2025]. Available at: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/correto/.

MICHAELIS
2025b *Dicionário Brasileiro da Língua Portuguesa*. [Online]. São Paulo: Melhoramentos, 2025. [Retrieved 15 Jan. 2025]. Available at: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/coerente/.

MIRANDA, Tânia Lúcia dos Santos
1995 *Estudos com a calicreína urinária humana: A - um novo método para purificação da enzima em larga escala, B - caracterização cinética com substratos sintéticos dos tipos amida e éster, derivados da arginina N-substituída e com os inibidores aprotinina e benzamidina*. [Online]. Belo Horizonte, 1995. [Retrieved 9 Jan. 2024]. Available at: http://hdl.handle.net/1843/BUOS-9NBKNE.
PhD thesis in Biochemistry and Immunology - Instituto de Ciências Biológicas, Universidade de Federal de Minas Gerais.

MOOERS, Calvin N.
1951 Zatocoding applied to mechanical organization of knowledge. *American Documentation*. [Online] 2:1 (1951) 20-32. [Retrieved 21 Nov. 2024]. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090020107.

OPENAI
2024a *About*. [Online]. 2024. [Retrieved 17 Dec. 2024]. Available at: https://openai.com/about/.

OPENAI
2024b *File Uploads FAQ*. [Online]. 2024. [Retrieved 17 Dec. 2024]. Available at: https://help.openai.com/en/articles/8555545-file-uploads-faq.

PATIL, Rajvardhan; GUDIVADA, Venkat
2024 A Review of current trends, techniques, and challenges in Large Language Models (LLMs). *Applied Sciences*. [Online]. 14:5 (2024). [Retrieved 1 Sept. 2024]. Available at: https://www.mdpi.com/2076-3417/14/5/2074.

ROSA, Flávia; GOMES, Maria João
2010 Comunicação científica: das restrições ao acesso livre. In *Repositórios institucionais: democratizando o acesso ao conhecimento*. Org. Maria João Gomes e Flávia Rosa. Salvador: EDUFBA, 2010, p. 11-34.

SARACEVIC, Tefko
1996 Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*. [Online]. 1:1 (1996) 41-62. [Retrieved 3 Aug. 2024]. Available at: https://periodicos.ufmg.br/index.php/pci/article/view/22308.

SHAHRIAR, Sakib; HAYAWI, Kadhim
2023 Let's have a chat!; A conversation with ChatGPT: Technology, applications, and limitations. *arXiv*. [Online]. 2302.13817v4 (2023). [Retrieved 17 Dec. 2024]. Available at: https://arxiv.org/abs/2302.13817?utm_source=chatgpt.com.

SOUZA, Rodrigo Ananias da Silva; RODAS, Cecílio Merlotti
2020 Recuperação da informação em dispositivos móveis. *Biblos: Revista do Instituto de Ciências Humanas e da Informação*. [Online]. 34:2 (2020) 147-166. [Retrieved 9 Jan. 2025]. Available at: https://doi.org/10.14295/biblos.v34i2.11840.

STATISTA
2024 *Volume of data/information created, captured, copied, and consumed world wide from 2010 to 2023, with forecasts from 2024 to 2028: in zettabytes*. [Online]. New York: Statista, 2025. [Retrieved 2 Sept. 2025]. Available at: https://www.statista.com/statistics/871513/worldwide-data-created/.

STOCK, Wolfgang G.; STOCK, Mechtild
2013 *Handbook of Information Science*. Berlin: De Gruyter, 2013.

TOPOL, Eric
2024 *Medicina profunda, deep medicine: como a inteligência artificial pode reumanizar os cuidados de saúde*. Porto Alegre: Artmed, 2024.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
2024a *UFMG em rankings*. [Online]. Belo Horizonte, 2024. [Retrieved 29 Aug. 2024]. Available at: https://ufmg.br/a-universidade/apresentacao/ufmg-em-rankings.

UNIVERSIDADE FEDERAL DE MINAS GERAIS. Repositório Institucional
2024b *Formulário de contato do RI-UFMG: Dúvida: Comunidade trabalhos acadêmicos, teses, dissertações e TCC digitalizadas, To: campos-daiane@ufmg.br*. Belo Horizonte, 11 Dec. 2024. Electronic message.

VAJJALA, Sowmya [et al.]
2020 *Practical Natural Language Processing: A Comprehensive guide to building real--world NLP systems*. Sebastapol, CA: O'Reilly, 2020.

WEI, Wendy Ran; HUANG, Ling; WANG, Jay Jianqiang
2025 *Retrieval-Augmented Generation for LLM applications: transforming search, recommendation, and AI assistants*. Sebastopol, CA: O'Reilly, 2025.

Daiane Campos Procópio | campos-daiane@ufmg.br
Escola de Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Brasil

Patrícia Nascimento Silva | patricians.prof@gmail.com
Escola de Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Brasil

Renato Rocha Souza | rsouza.fgv@gmail.com
Escola de Ciência da Informação, Universidade Federal de Minas Gerais (UFMG), Brasil

Appendix
Record of the analysis of the responses generated by the model

Identification		Quantitative analysis	Qualitative analysis	
Thesis ID	**Prompt No.**	**Did the model generate a correct response?**	**Did the model generate a coherent response?**	**Likert Scale assigned**
1	1	Yes	Yes	5
1	2	Yes	Yes	5
1	3	Yes	Yes	5
1	4	Yes	Yes	5
1	5	Yes	Yes	5
6	1	Yes	Yes	5
6	2	Yes	Yes	5
6	3	Yes	Yes	5
6	4	Yes	Yes	5
6	5	Yes	Yes	5
9	1	Yes	Yes	5
9	2	Yes	Yes	5
9	3	Yes	Yes	5
9	4	Yes	Yes	5
9	5	Yes	Yes	5
10	1	Yes	Yes	5
10	2	Yes	Yes	5
10	3	Yes	Yes	5
10	4	Yes	Yes	5
10	5	Yes	Yes	5
11	1	Yes	Yes	5
11	2	Yes	Yes	5
11	3	Yes	Yes	5
11	4	Yes	Yes	5
11	5	Yes	Yes	5
12	1	Yes	Yes	5
12	2	Yes	Yes	5
12	3	Yes	Yes	5
12	4	Yes	Yes	5
12	5	Yes	Yes	5
14	1	Yes	Yes	5
14	2	Yes	Yes	5
14	3	Yes	Yes	5
14	4	Yes	Yes	5
14	5	Yes	Yes	5
15	1	Yes	Yes	5
15	2	Yes	Yes	5
15	3	Yes	Yes	5
15	4	Yes	Yes	5
15	5	Yes	Yes	5
16	1	Yes	Yes	5
16	2	Yes	Yes	5
16	3	Yes	Yes	5
16	4	Yes	Yes	5
16	5	Yes	Yes	5
17	1	No	No	1
17	2	Yes	Yes	5
17	3	No	No	1
17	4	Yes	Yes	5
17	5	Yes	Yes	5
18	1	Yes	Yes	5
18	2	Yes	Yes	5

18	3	Yes	Yes	5
18	4	Yes	Yes	5
18	5	Yes	Yes	5
19	1	Yes	Yes	5
19	2	Yes	Yes	5
19	3	Yes	Yes	5
19	4	Yes	Yes	5
19	5	Yes	Yes	5
20	1	Yes	Yes	5
20	2	Yes	Yes	5
20	3	Yes	Yes	5
20	4	Yes	Yes	5
20	5	Yes	Yes	5
21	1	Yes	Yes	5
21	2	Yes	Yes	5
21	3	Yes	Yes	5
21	4	Yes	Yes	5
21	5	Yes	Yes	5
22	1	Yes	Yes	5
22	2	Yes	Yes	5
22	3	Yes	Yes	5
22	4	Yes	Yes	5
22	5	Yes	Yes	5
23	1	Yes	Yes	5
23	2	Yes	Yes	5
23	3	Yes	Yes	5
23	4	Yes	Yes	5
23	5	Yes	Yes	5
24	1	Yes	Yes	5
24	2	Yes	Yes	5
24	3	Yes	Yes	5
24	4	Yes	Yes	5
24	5	Yes	Yes	5
25	1	Yes	Yes	5
25	2	Yes	Yes	5
25	3	Yes	Yes	5
25	4	Yes	Yes	5
25	5	Yes	Yes	5
26	1	Yes	Yes	5
26	2	Yes	Yes	5
26	3	Yes	Yes	5
26	4	Yes	Yes	5
26	5	Yes	Yes	5
27	1	Yes	Yes	5
27	2	Yes	Yes	5
27	3	Yes	Yes	5
27	4	Yes	Yes	5
27	5	Yes	Yes	5