Information Technology & Tourism
https://doi.org/10.1007/s40558-025-00339-x

CASE STUDY

Lisa: a touristic chatbot for Lisbon

Miguel Cruz[1] · **Bruno Jardim**[1] · **Miguel de Castro Neto**[1]

Received: 21 January 2025 / Revised: 12 July 2025 / Accepted: 7 August 2025

Abstract

As cities continue to attract a growing number of visitors, the development of a tailored chatbot catering to the tourists' unique needs becomes increasingly valuable. In this paper, we propose a new methodology for developing generative implementations of touristic chatbots and demonstrate its application through a prototype created specifically for Lisbon. Utilizing a web-scraped knowledge base with over 2000 website pages, the chatbot offers recommendations for tourist routes, events and places to visit, provides general information about Lisbon (including transportation) and engages with visitors. The initial evaluation was conducted using synthetic datasets—one for question-answering and another for recommendations—to guide experiments in data preprocessing, exploration of different ChatGPT models and improvements to the Retrieval-Augmented Generation pipeline. A subsequent evaluation was performed with experts, including professionals from Turismo de Portugal, focusing on three major areas (recommendation, information and engagement), and the chatbot achieved very strong results on all use cases. This paper contributes to the literature on chatbot development, emphasizing the benefits of advanced machine learning models in the tourism industry and highlighting the potential of iterative optimization and evaluation based on both synthetic and human survey data for downstream tasks. Likewise, the proposed approach can significantly enhance the tourism experience by offering personalized, engaging and efficient digital interactions, thereby improving overall visitor satisfaction and supporting sustainable destination management.

Keywords Chatbot · Transformer · Chatgpt · Tourism · Retrieval augmented generation

Extended author information available on the last page of the article

Published online: 14 August 2025

1 Introduction

Lisbon, Portugal, has established itself as a renowned tourist destination, with a consistent trend of growing popularity over time. In fact, in 2023, the city hosted 8,819,500 visitors (TravelBI by Turismo de Portugal 2024). Its historical significance, cultural offerings, and scenic vistas collectively make Lisbon an appealing destination for visitors interested in history, culture, and natural beauty. Concurrently, Information and communication technologies have reshaped the tourism landscape, offering improved services and heightened community engagement (Anand et al. 2023). Notably, chatbot technology has emerged as an important tool (Calvaresi et al. 2021). This technology bridges the gap in catering to the growing user needs for information, recommendation and engagement, enhancing the overall tourist experience.

As visitor numbers to Lisbon increase, there is a stronger demand for these tools. They efficiently handle customer inquiries, freeing up human staff for personalized service and improving overall visitor satisfaction. With their 24/7 availability, chatbots address diverse tourist needs and mitigate staffing challenges (Bulchand-Gidumal et al. 2024). They can also provide a more dynamic and interactive way of delivering information (Calvaresi et al. 2021), fostering meaningful engagement with users. Financially, chatbots save time, automate tasks and enhance sales, making them essential in modern tourism management (Calvaresi et al. 2021). Therefore, there is strong value in developing a chatbot to address the unique needs of Lisbon's tourists, ranging from local event recommendations to mobility questions. Leveraging current advances in Machine Learning (ML), chatbots can now offer sophisticated language understanding and generation capabilities (Caldarini et al. 2022).

Still, there is a scarcity of literature focusing on chatbots utilizing advanced Artificial Intelligence (AI) capabilities, particularly leveraging the transformer architecture (Caldarini et al. 2022) and harnessing the potential of models like ChatGPT, especially in the tourism industry. While some studies in the tourism industry may indirectly utilize the transformer architecture through third-party applications, no explicit mention of its use has been found. Moreover, persistent obstacles—such as difficulty understanding complex queries, maintaining coherent, natural dialogue and generating contextually relevant recommendations—continue to limit chatbots' practical value for the industry (Calvaresi et al. 2021). Addressing these gaps can be achieved through more advanced, transformer-based chatbots.

Thus, in this study, we propose a new methodology for developing generative implementations of touristic chatbots, using Lisbon as a prototype city. The goal is to create an engaging general-purpose chatbot that can fulfil the unique needs of Lisbon's tourists. This involves recommending places to visit, tourist routes and local events, providing general information about Lisbon and engaging with visitors. Through the chatbot, we can address the specific requirements of tourists in Lisbon, enhancing their satisfaction and fostering the growth of tourism in the city.

This methodology for developing generative touristic chatbots leverages a web-scraped knowledge base of over 2,000 pages, integrated into a Retrieval-Augmented Generation (RAG) pipeline with optimized prompts and state-of-the-art transformer architectures (ChatGPT) to deliver a sophisticated conversational system. Experi-

ments involve data preprocessing, exploration of different ChatGPT models and improvements to the RAG pipeline. Evaluation follows a two-stage process. In the first stage, two synthetic datasets were created: one for evaluating recommendation capabilities, assessing the chatbot's ability to respond to typical open-ended user queries, and another for question-answering, testing its ability to provide accurate responses based on the knowledge base. In the second stage, three experts, including professionals from Turismo de Portugal, evaluated the final version of the chatbot across three key areas: recommendation, information and engagement.
With this, the paper makes the following contributions to the literature on chatbot development:

- Address critical limitations in tourism chatbots' language capabilities through a large language model based architecture implementation, specifically using ChatGPT.
- Demonstrate a systematic and scalable optimization methodology that iteratively measures the performance of each of the system components.
- Develop an evaluation methodology that combines rapid evaluation for system components through synthetic data with human assessment for more detailed feedback closer to the application scenario.
- Show how advanced chatbots can enable contextually-aware recommendations, natural booking conversations, and data-driven destination management insights through a prototype and its evaluation results.

The remainder of this study proceeds as follows. In Sect. 2, we describe the technical background behind the development of chatbots and present previously implemented city tourism chatbots, along with relevant examples of models using the Transformers architecture. In Sect. 3, we detail the model architecture, outline the experiments and the human evaluation conducted. Section 4 displays and analyses the results of the evaluations. Section 5 discusses the findings and their implications. Section 6 concludes the work, adding possible ideas to be developed in future research.

2 Literature review

Chatbots are software applications designed for simulating conversations with human users. Usually, they are given as input text and respond with relevant output sentences (Caldarini et al. 2022). In academic and research contexts, these systems can also be referred as intelligent agents, virtual assistants, dialogue systems and conversational agents (Kusal et al. 2022).

Chatbots started with Alan Turing's 'Turing Test', which was designed to evaluate a machine's ability to mimic human-like conversation (Turing 1950) and became a foundational benchmark in AI. Afterwards, significant milestones were achieved with the creation of ELIZA and Artificial Linguistic Internet Computer Entity (ALICE). ELIZA, developed by Weizenbaum (1966) at MIT, is often considered the first chatbot. It used pattern matching to simulate conversation, especially in the mode of a Rogerian psychotherapist and demonstrated the potential of simple rule-based algo-

rithms in generating human-like text interaction. Subsequently, ALICE created by Richard Wallace in 1995, marked an advancement in the field. ALICE utilized the Artificial Intelligence Markup Language (AIML), which is an extension of Extensible Markup Language (XML) and was instrumental in demonstrating the capabilities of chatbots for more varied and complex interactions (Caldarini et al. 2022).

With the progressive increase in computational power and advances in ML and Deep Learning (DL), the chatbots evolved from simple pattern matching to more sophisticated language understanding and generation (Caldarini et al. 2022). ML can be applied both in Natural Language Understanding (NLU), by understanding better a user's query, and in Natural Language Generation (NLG), by generating a response without the need for pre-defined answers. A pivotal development in this domain has been the introduction of transformer models. Unlike previous models that processed words in sequence, transformers are capable of processing entire blocks of text in parallel, allowing for more efficient and contextually nuanced language understanding (Vaswani et al. 2017). This has led to the creation of more advanced chatbots that can understand and generate human-like text. ChatGPT, a recent and significant development in the realm of chatbots, is based on a transformer architecture. It has gained considerable attention for its ability to generate coherent and realistic responses across diverse topics (Lund and Wang 2023).

Considering the novelty of many of the subject areas being explored, this section will start by presenting the technical background behind the creation of a chatbot. After this contextualization, related works will be presented.

2.1 Background

Chatbots vary significantly in terms of domain, implementation, evaluation and optimization methods. Each subsection below will address one of these aspects.

2.1.1 Open-domain and closed-domain chatbots

Open-domain chatbots are designed to engage in conversations on a wide range of topics, without being restricted to a specific domain or subject matter (Doğruöz and Skantze 2022). This flexibility allows them to simulate more natural, human-like interactions, as they can discuss anything from daily news to general knowledge. Nowadays, these chatbots often rely on large language models (LLMs), computational models that have the capability to understand and generate human language, which enable them to process and generate responses to diverse queries. A prime example is OpenAI's Generative Pre-trained Transformers (GPT) series, including models like ChatGPT. These chatbots are trained on extensive datasets encompassing a broad spectrum of human knowledge and interactions (Lund and Wang 2023). Evaluating open-domain chatbots can be difficult as it involves assessing their ability over a wide range of subjects and applications (Doğruöz and Skantze 2022).

In contrast, closed-domain chatbots are specialized tools designed for specific tasks or subject areas. They excel in environments where expertise or focused assistance is required, such as customer service or educational support. These chatbots typically use a narrower dataset relevant to their specific domain, and their algo-

rithms are optimized for understanding and responding to queries within that context. The evaluation of closed-domain chatbots often focuses on their ability to execute domain-specific tasks (Lin et al. 2023).

2.1.2 Rule-based and AI-based implementations

Irrespective of their domain, chatbots can have rule-based or AI-based implementations. Rule-based chatbots rely on predetermined rules and pattern matching to respond to user queries. These models are easier to design and implement but struggle with complexity and ambiguity in queries, often resulting in inaccurate answers outside their programmed patterns. Additionally, pattern matching rules are highly domain-specific, time-consuming to encode and lack adaptability across various problems (Caldarini et al. 2022). The first tries at creating a chatbot were rule-based (both ELIZA and ALICE).

AI-based chatbots leverage ML algorithms to respond to user queries, learning from data containing human conversations. Using ML, there is no need to define pattern matching rules, allowing for more flexibility and adaptability across distinct domains, and better suited to handling complex and ambiguous textual data scenarios. AI-based models can be divided into Information Retrieval (IR) based models and Generative Models (Caldarini et al. 2022).

IR models focus on retrieving the relevant answer based on the user's query from an existing dataset. These models excel at matching user queries with high-quality answers (Caldarini et al. 2022). However, creating this dataset, with all the predefined answers, can be very costly and time-consuming. As well as that, it suffers from domain dependency (Kusal et al. 2022).

Generative AI models, such as ChatGPT, operate by generating responses based on input data, learning from vast datasets to understand language nuances, context and conversational patterns. Utilizing DL techniques and neural network architectures, most notably encoder-decoder, these models can produce human-like answers (Lund and Wang 2023).

A prime example of this encoder-decoder approach is the Transformer architecture, which has revolutionized natural language processing, surpassing models such as convolutional and recurrent neural networks in performance for tasks in both natural language understanding and natural language generation (Wolf et al. 2019).

ChatGPT, a Large Language Model (LLM) developed by OpenAI, is built upon a transformer architecture, specifically the Generative Pre-trained Transformer (GPT) fine-tuned for conversational purposes, marking a significant milestone in the adoption of transformer-based models. The technology's wide-ranging implications are notable in both scientific and societal contexts, with potential impacts across various industries and fields (Fraiwan and Khasawneh 2023). ChatGPT-3 uses 175 billion parameters and was trained on a dataset of 300 billion words to accurately answer user queries (Kenney 2023). ChatGPT-4 presented a substantial leap forward, with 1 trillion parameters and human-level performance on many professional and academic benchmarks (Mijwil et al. 2023; OpenAI et al. 2023). Additionally, the GPT-4o architecture matches GPT-4 Turbo performance on text in English and code, with

significant improvement on text in non-English languages, while also being much faster and cheaper (OpenAI et al. 2024).

2.1.3 Evaluation

After the definition of the domain and the implementation of the chatbot, there is a need to evaluate its performance. Evaluation of chatbots involves two primary methods: automatic and human evaluations. These approaches serve to assess the performance and capabilities in different ways. Automatic evaluation relies on standardized metrics and evaluation tools to measure model performance. In contrast, human evaluation involves human judgment and participation to assess the quality of model-generated outputs. These methods complement each other, providing distinct perspectives on the model's effectiveness (Chang et al. 2023).

Automated evaluation metrics offer efficiency in terms of time and resources required for assessment. However, there remains a lack of established industry standards concerning the application of these metrics. While automated evaluation metrics are easily implementable, they seem inadequate in comprehensively assessing the overall quality, efficiency, and effectiveness of conversational systems (Caldarini et al. 2022).

Human evaluation in the context of chatbots involves inviting participants to interact with the chatbot and then assess various factors using pre-defined frameworks or questionnaires. Despite providing insights into diverse interaction dimensions, human evaluation possesses certain drawbacks. It tends to be costly, time-consuming, and not easily scalable (Mnasri 2019). However, it aligns more closely with the actual application scenario, providing accurate results (Chang et al. 2023).

2.1.4 Methods to optimize LLM performance

With an evaluation framework, optimization is possible as you can identify which experiment is performing the best. In fact, when transitioning from prototype to production, optimizing a model becomes necessary. As a LLM (ChatGPT) is used in this paper, emphasis will be placed on optimization techniques tailored for LLMs. Commonly employed techniques include prompt engineering, RAG and fine-tuning.

Prompt engineering is the process of designing and optimizing the input prompts to guide the output of the LLM. This can help improve the versatility and relevance of the LLMs answers and counteract machine hallucinations (coherent outputs but factually incorrect). This should be the initial focus of the optimization, given its relatively simpler implementation. Numerous approaches exist for this purpose. Key elements of an effective prompt involve providing clear and precise instructions, employing role-prompting and utilizing one-shot or few-shot prompts, by providing one or multiple examples (Chen et al. 2023).

RAG further enhances the accuracy and contextual relevance of LLM outputs by merging parametric memory learned during training with non-parametric memory sourced from an external knowledge base (Lewis et al. 2020). This knowledge base consists of task-specific documents, such as information about mobility in Lisbon, embedded as high-dimensional vectors and usually stored in vector databases. Vector

databases provide fast and accurate similarity search and retrieval and support for complex and unstructured data, with a small trade off on accuracy (Han et al. 2023). When the user writes its query, the model retrieves the most important information from this knowledge source and augments the query with this additional context. Thanks to the provided context and this new engineered prompt, the model is able to generate accurate and factual responses. Different vector databases, retrieval algorithms or augmentation mechanisms can impact the performance of this system.

Fine-tuning involves modifying the model's parameters using a smaller, task-specific dataset. It differs from prompt engineering and RAG as they solely alter the prompt. According to OpenAI (2023), fine-tuning is effective for emphasizing existing knowledge within the model, customizing response structure or tone and teaching a model highly complex instructions. However, it is not good for incorporating new knowledge into the base model (RAG is more suitable for this purpose) or rapidly iterating on a new use case due to the substantial effort it requires compared to other alternatives.

OpenAI (2023) structured the optimization flow using two axes: Context optimization (what the model needs to know) and LLM optimization (how the model needs to act), as can be seen in the Fig. 1.

Prompt engineering serves as an initial solution that can enhance both the model's behaviour and its existing knowledge. RAG is particularly advantageous for refining the model's knowledge (optimizing its context), whereas fine-tuning aids in instructing the model with established behaviours. In certain use cases, all these solutions can be employed simultaneously.

2.2 Related works

In this section, we will begin by examining chatbots tailored for tourism (subsection 2.2.1). Given the limited research on transformer architecture and ChatGPT in this field, we will explore examples from other industries (subsection 2.2.2).

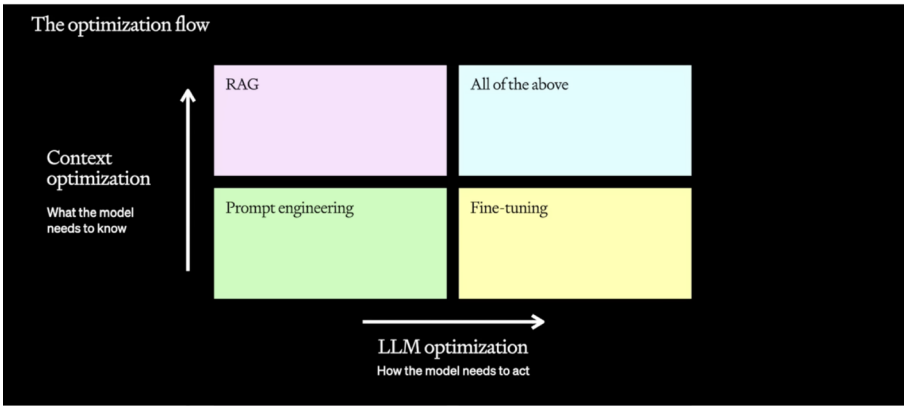

Fig. 1 The optimization flow for an LLM. From *A Survey of Techniques for Maximizing LLM Performance*., by OpenAI (2023)., https://www.youtube.com/watch?v=ahnGLM-RC1Y

2.2.1 Touristic chatbots

Most chatbots in the industry utilize rule-based interactions, such as standardized menus, avoiding the need for complex language processing. While these models are easier to design and develop, their capabilities are limited. The rules they rely on must be highly domain-specific, making them unable to adapt to broader contexts and complex queries. As a result, rule-based models are typically satisfactory only in specific scenarios (Caldarini et al. 2022).

Casillo et al. (2020) developed a rule-based chatbot prototype as a recommendation system for the Campania region in Southern Italy. The system, designed to provide adaptive tourist routes, used pattern and context recognition techniques to suggest points of service and related services based on the tourist's profile and contextual aspects. Although feedback from 3150 users was largely positive, the lowest ratings were in the conversation category, where users encountered difficulties in maintaining a fluent dialogue with the chatbot and experienced challenges in having their requests accurately understood. The Rahhal chatbot, developed by Alhumoud et al. (2022) further illustrates the limitations of rule-based systems in tourism. Like the Italian example, Rahhal functions primarily as a recommendation engine for tourists in Saudi Arabia, helping users find activities and interesting locations in 81 main cities. While useful for this specific purpose, the chatbot's capabilities are restricted to providing recommendations, lacking the versatility to address other crucial aspects of tourism such as disseminating general information, handling complex queries or offering a more comprehensive conversational experience. Another instance of this type of implementation is Amalia and Suprayogi (2019) work about the development of a rule-based chatbot in Yogyakarta—accessible via Telegram and Facebook Messenger—supplemented by an informative website to promote eco-tourism. The chatbot provided users with information about eco-tourism and eco-tourism clusters in Yogyakarta, drawing data from sources including crowd-sourcing-based search engines and verified websites. The chatbot implementation relies on a relatively simple rule-based structure, with menu-based queries.

There have been studies trying to tackle these challenges through Information Retrieval and Generative approaches. Anand et al. (2023) developed a chatbot for Indian cities using a Convolutional Neural Network (CNN) with TensorFlow. The CNN allowed users to input queries in natural language, which the system then matched to pre-defined answers to provide real-time information on service availability, such as hotels and hospitals. This approach enabled more flexible user interactions while still relying on a retrieval-based method for responses.

D. Arteaga et al. (2019) developed a chatbot with IBM Watson Assistant to present personalized recommendations of tourist sites in Manta, Ecuador, via Facebook Messenger and an informative website. The system combined Watson's conversational interface with a recommendation engine, illustrating that delivering personalized suggestions is a compelling use case for chatbots and pointing toward the promise of more advanced AI-based chatbot development. Sperlí (2021) also proposed a more advanced framework for supporting tourists' journeys, using a generative approach. The framework employs a Seq2Seq model with Gated Recurrent Unit (GRU) cells for its chatbot engine, demonstrating significant improvements over traditional rule-

based systems. The experimental results showed that Sperlí's approach outperformed state-of-the-art models in terms of efficiency and efficacy, including Casillo et al. (2020) chatbot previously mentioned.

However, significant challenges remain. The inability to process complex information, provide natural interactions and generate appropriate recommendations continues to hinder the widespread adoption of chatbots in the industry (Calvaresi et al. 2021). Sperlí's results highlight the promise of generative approaches in touristic chatbots, but further advancements are still needed. The introduction of transformer-based architectures offers a promising direction in the development of more sophisticated and responsive conversational systems.

2.2.2 Models using the transformer architecture

Other industries have benefited from the use of the transformer architecture. Its success across domains such as education and disaster management highlight its versatility and potential to address the limitations found in tourism chatbots.

Nhut Lam et al. (2023) proposed an educational virtual assistant consisting of two integrated chatbots, both using transformers: a closed-domain chatbot, trained on over thirty-five thousand factual question-answer pairs to engage in university-related conversation and an open-domain chatbot, trained on a large movie dialog dataset to engage in general conversation. The objective is to support students in Vietnamese. In experiments with real users, the system successfully answered questions outside its training dataset, understood user queries with misspellings or missing diacritics and provided contextually relevant and accurate responses.

Boné et al. (2020) developed a Portuguese-speaking chatbot designed to support disaster management. The chatbot was developed using the RASA framework, incorporating the Dual Intent Entity Transformer (DIET) architecture for user intent classification. DisBot integrates real-time updates from a dynamic knowledge graph, enabling it to provide timely and relevant information. Its outputs include both predefined answers for general conversation and specific, knowledge-based responses to disaster-related inquiries. Examples of the first include chitchat to maintain user engagement and questions to gather necessary details from the user. Knowledge-based responses are generated from queries to the dynamic knowledge graph and provide users with details on the disaster situation and recommended actions. Validation by field specialists confirmed its strong performance, achieving from Largely achieved to Fully Achieved on all proposed criteria.

Transformers have the capacity to address more complex and dynamic use cases. The advanced natural language understanding and generation capabilities of ChatGPT (OpenAI et al. 2023), in particular, provide a unique opportunity to address the limitations of current touristic chatbots. However, to deploy LLMs like ChatGPT effectively in production environments, optimization is critical to improve performance and align the models with specific tasks. While the field of optimizing ChatGPT for production use is still emerging, recent research highlights prompt engineering, fine-tuning and RAG as promising methods.

Li et al. (2023) used fine-tuning and prompt engineering (role-prompting with clear and precise instructions) to implement ChatGPT-3.5 Turbo as a news recom-

mendation system. The methodology involved fine-tuning ChatGPT on the Microsoft News dataset (MIND) by designing two types of prompts: one for a ranking task (where the model orders candidate articles based on user reading history) and another for a rating task (where the model predicts a rating score for each candidate article). The model showed competitive results to the alternative methods (NAML, LSTUR, NRMS and Popularity), even surpassing them in ranking tasks. Fine-tuning proved effective in enhancing the model's understanding of news and recommendations, as well as ensuring structured output in accordance with specified formatting criteria, while prompt engineering was used to accurately establish the initial context.

Silva et al. (2023) demonstrated the effectiveness of RAG and prompt engineering by evaluating popular LLMs, such as Llama 2 and GPT-3.5, on agriculture-related queries. The data used to provide context for the LLMs via RAG consisted of the text materials given to certified agronomists for their continuing education requirements. With this, the LLM had access to the relevant, external knowledge needed for each of the queries. Their study showed that incorporating RAG significantly enhanced model performance in domain-specific question-answering tasks. By applying both RAG and prompt engineering to ChatGPT-4, they achieved a 13-percentage-point improvement in accuracy for AgriExams questions.

The research and development of production-ready RAG systems are still in their early stages (Barnett et al. 2024). Still, it is evident that RAG and ChatGPT, in general, can be enhanced and optimized.

3 Methods

3.1 Model architecture

This study utilizes two chatbot models: ChatGPT-3.5 Turbo with a 16 K context window and ChatGPT-4o with a 128 K context window. Context window refers to the maximum number of tokens per input request. When the context window is exceeded, the oldest previous answers are removed until the window size is within limits again. The OpenAI library's ChatCompletion.create() function was employed and it works by receiving a message, in natural language, with the specified roles (e.g. system and/or user) and returning an answer, also in natural language. The system message helps define the behavior of the assistant across all interactions, while the user message provides questions or comments for the assistant to answer.

Figure 2 illustrates the process starting from the user message to a corresponding response, assuming a vector database is being used.

The figure demonstrates the RAG workflow used, where the chatbot leverages a knowledge base to generate accurate and context-aware responses. When a user inputs a query, such as asking for recommendations about Lisbon, the system first performs a retrieval step to obtain relevant contextual data (e.g. document about Palácio Nacional de Belém). This retrieved data is then combined with the user's input and system prompt to create an augmented prompt, optimizing the input for the chatbot. Finally, the chatbot uses this augmented prompt to generate a response that is both relevant and informative, addressing the user's request effectively. For our

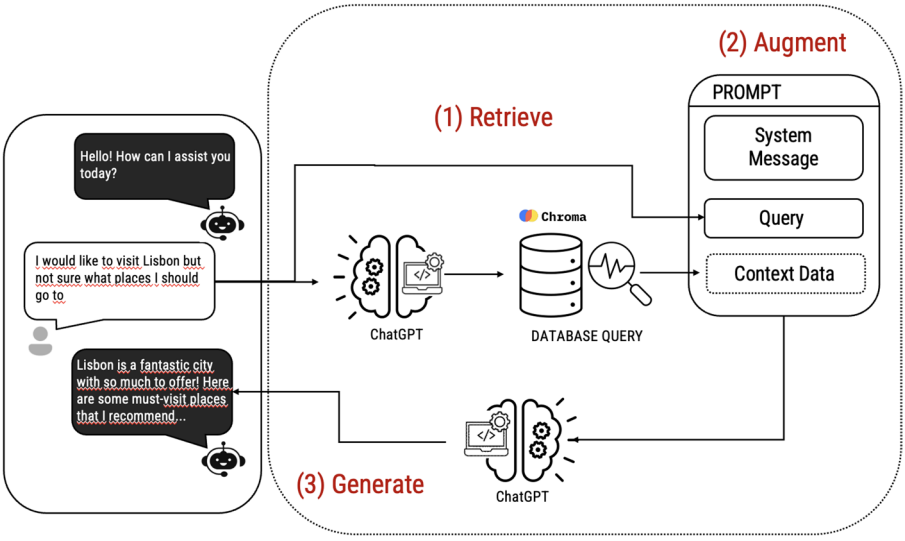

Fig. 2 Example of Lisa's retrieval-augmented generation process from user message to answer

study, we used Chroma (Chroma, 2024) as the vector database due to its open-source nature.

The knowledge base used was gathered between December 2023 and January 2024, primarily sourced from the Visit Lisboa website (Visit Lisboa, 2024), a governmental platform dedicated to providing information for tourists visiting Lisbon, as described in Table 1.

A part of the data is in Portuguese, as the websites did not provide a translated version. Automated data collection was performed using Selenium's web driver (Selenium, 2024), which programmatically navigated the web pages to extract the required information. However, some data had to be manually written or copy-pasted because it was either not in an easily extractable format (e.g. images) or was too minimal to justify automation.

Certain preprocessing steps were implemented to ensure the quality of the knowledge base. These included removing duplicate and redundant website addresses, correcting inconsistencies in TripAdvisor review links and addressing documents flagged for re-execution. Invalid links and inappropriate content were identified and removed during token analysis and chatbot testing. To manage content exceeding the token limit, repetitive elements and recommendations from TripAdvisor pages were truncated. Additionally, any information violating Azure OpenAI's content management policy was eliminated.

The system prompt was designed following the foundational principles of Prompt Engineering (Chen et al. 2023): role-prompting (AI assistant for tourists), providing clear and precise instructions (how it should answer and what it will have access to) and having an example. The prompt included advice to exercise caution when responding to queries about pricing. Despite the availability of pricing data in the knowledge base, this information should be handled carefully as it may not be updated or entirely accurate, being sourced from a third-party like TripAdvisor. As

Table 1 Data description of knowledge base

Source	Description	Method	Language
Visit Lisboa	General information about Lisbon	Automated	English
Visit Lisboa	Lisbon stories (including the description of many important cultural elements)	Automated	English
Visit Lisboa	More than 500 experiences (including events, restaurants, hotels and others)	Automated	English
Visit Lisboa	Lisboa card details	Manual	English
TripAdvisor	Pages associated to each of the experiences in Visit Lisboa	Automated	English
Agenda Cultural de Lisboa	More than 500 additional events (to supplement Visit Lisboa events' listings)	Automated	Portuguese
Carris	Useful transportation information	Automated	English
Comboios de Portugal	Useful transportation information	Automated	English
Metro Lisboa	Useful transportation information	Automated	English
Metro Lisboa	Information about the different metro lines	Manual	English
Metro Lisboa	Art featured in the metro stations	Automated	Portuguese
Reddit	Problems in traffic, including best and worst times to leave	Manual	English
None	Added sentence indicating where to buy tickets and reserve accommodations	Manual	English

both models do not have access to current date and time, a sentence was added to the prompt to programmatically indicate the time. The user prompt corresponds to the actual query with the added context and it also includes the previous queries and answers in the conversation.

Both the system and user prompts were refined through iterative testing, with manual review of the model's responses guiding each adjustment to tone, instruction clarity and contextual framing. This prompt-engineering approach enabled rapid experimentation without retraining the model, allowing us to focus immediately on how it interpreted and acted on our guidance. For these early stages, we chose prompt engineering over fine-tuning because it demanded fewer computational resources, shorter development cycles and minimal setup, while still delivering the desired conversational style. As our primary challenge lay in equipping the model with up-to-date, domain-specific information, we determined that building a retrieval-augmented knowledge base was a higher priority than investing time and effort into full model fine-tuning. After establishing the prompts with the desired initial context, instruction definition and tone, we recognized that running formal prompt-variation experiments would both increase computational overhead and present challenges in either automating tone analysis or managing repeated human evaluations. Because the prompts already achieved the desired tone and produced responses that were accurate with respect to the given context and instructions, we did not systematize those tests for the purposes of the study.

3.2 Experiments

Experiments were conducted to enhance the chatbot's performance and response time by systematically analyzing the impact of key system components. The goal was to determine how changes to the vector database, retrieval algorithm, similarity measure, preprocessing methods, number of retrieved documents, temperature parameter and ChatGPT model influence the system's results.

We started by testing the vector database (Chroma) because it could improve significantly the speed of the remaining experiments. Next, the retrieval algorithm and the datasets forming the knowledge base were analyzed, as these are critical to the accuracy of the RAG workflow. Additional experiments focused on optimizing system parameters, such as the number of retrieved documents and the temperature, which, while smaller components, could still contribute to better performance. Finally, the experiments included an evaluation of the ChatGPT model, particularly ChatGPT-4o, which was tested later due to its recent release.

These experiments contribute to the development of a reliable and scalable methodology for generative chatbot implementations in tourism. By systematically testing and refining key components, we ensure the development of a robust system, designed to be adaptable for further domain-specific applications and ready for expert evaluation.

3.2.1 Retrieval algorithms

Three different retrieval algorithms were tested, based on distinct methods: algorithm based on OpenAI embeddings, BM25 and Hybrid Search, combining the set of ranked documents of the first two. In addition to exploring various retrieval algorithms, Chroma, an open-source embedding database as mentioned previously, was utilized to enhance retrieval speed by storing the embedded documents from the knowledge base and providing fast retrieval capabilities. The decision on the number of documents retrieved was made in accordance with the maximum context window, following further preprocessing.

The text embedding-ada-002 model from OpenAI is one of the strongest multilingual embeddings, as can be seen from the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al. 2022). This model was used to obtain separate embeddings for the query and each document in the knowledge base. Then, the similarity between the embeddings of the query and each document was calculated, retrieving those documents with the highest similarity (a simple for loop was used). The two similarity measures used in this research were: squared L2 norm and cosine similarity. These measures were employed because both are prevalent choices utilized in textual similarity problems, as noted by Thompson et al. (2015).

BM25 was also utilized, as it is arguably one of the most important and widely used information retrieval functions (Svore and Burges 2009). The ranking scores to determine the relevance of documents to a certain search query are calculated as seen in (1).

$$rsv_q = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) . \frac{(k_1 + 1) . tf_{td}}{k_1 . \left(1 - b + b . \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \tag{1}$$

For a given query, q, the retrieval status value, rsv_q, is the sum of individual term, t, scores. N is the number of documents in the collections, df_t is the number of documents containing the term (the document frequency), tf_{td} is the number of times term t occurs in document d. L_d is the length of the document (in terms) and L_{avg} is the mean of the document lengths. There are two tuning parameters, b, and k_1 (Trotman et al. 2014).

The final approach, known as Hybrid Search, was based on the Ensemble Retriever by LangChain (2024). A sparse retriever, BM25, was combined with a dense retriever, algorithm based on OpenAI Embeddings, to improve the results. Each retriever was given equal importance in this paper. The sparse retriever is good at identifying relevant documents using keywords, while the dense retriever is adept at finding relevant documents based on semantic similarity.

Initially, each retriever independently generates a ranked list of documents based on its respective retrieval method. These ranked lists are then merged using the Reciprocal Rank Fusion (RRF) algorithm. RRF sorts the documents according to a naive scoring formula. Given a set D of documents and a set of rankings R, each a permutation on $1..|D|$, the score is determined as follows in (2).

$$\text{RRFscore}\,(d \in D) = \sum_{r \in R} \frac{1}{k + r\,(d)} \tag{2}$$

where k = 60 was fixed during a pilot investigation and not altered during subsequent validation (Cormack et al. 2009). For the experiments, 10 documents were ranked for each of the retrievers used.

3.2.2 Additional preprocessing

Additional preprocessing of the knowledge base was tested to refine the retrieval process and enhance the chatbot's capability to generate context-aware responses without being overwhelmed by unnecessary information. Initially, the dataset (refer to subsection 3.1) underwent basic preprocessing. Subsequently, two additional datasets were created by applying progressively refined preprocessing steps, resulting in three distinct datasets for testing.

Although it is uncommon to test different preprocessed knowledge bases during the development of the system, we considered this a critical factor influencing the chatbot's performance. Since the knowledge base forms the foundation for the chatbot's responses, we aimed to determine whether refining the data format could deliver better results. This approach was particularly relevant given the unstructured and inconsistent nature of web-scraped data.
For the first dataset, the following modifications were implemented:

- Elimination of TripAdvisor reviews for each experience, retaining only those

highlighted by the website itself.
- Removal of repetitive structural elements present across all metro pages.

For the second dataset, additional changes were made:

- Removal of repetitive structural elements shared across pages from Visit Lisboa, Agenda Cultural de Lisboa, Comboios de Portugal, Carris and TripAdvisor.
- Extraction of essential elements common to all pages into a separate text file, thereby eliminating them from the content of individual pages (e.g. TripAdvisor Traveler's Choice).
- Removal of details about hotel deals from TripAdvisor reviews.

3.2.3 Temperature

The temperature parameter for the chatbot model ranges from 0 to 1. Higher values increase randomness in the output, while lower values enhance focus and determinism (OpenAI, 2024). For instance, a chatbot designed for creative tasks would typically have a temperature closer to 1, whereas a chatbot focused solely on question-answering, where engagement is less critical, would have a temperature closer to 0.

In our case, we wanted to evaluate whether the temperature parameter significantly impacted the chatbot's performance. While the primary goal is to ensure the chatbot can effectively answer questions, it also needs to engage users with creative and dynamic responses. We limited our evaluation to three temperature values, as they yielded comparable performance (as will be presented in the following section).

3.2.4 ChatGPT model

Due to the recency of the release of ChatGPT-4o, most experiments were conducted with ChatGPT-3.5 Turbo. In the end, their performance was compared for only two scenarios: the baseline case (the simplest retrieval algorithm – algorithm based on OpenAI Embeddings – with the default similarity measure – L2, using 3 documents for ChatGPT-3.5 Turbo and 2 for ChatGPT-4o, the initial dataset, a temperature parameter of 0.5 and Chroma as a vector database) and the case with the best results for ChatGPT-3.5 Turbo. Due to strict rate limits in the Azure platform, the number of documents used for the ChatGPT-4o model were reduced compared to ChatGPT-3.5 Turbo.

3.2.5 Overview

All the performed experiments are outlined in Table 2. Each technique described above will be evaluated in sequence, with the optimal configuration moving forward to subsequent experiments. An exhaustive exploration of every possible permutation would undermine the methodology's scalability.

Table 2 List of experiments performed

Order	Chroma	Retrieval Algorithm	Dataset	Documents	Temperature	Model
1	Without	OpenAI Embeddings w/ L2	Initial	3	0.5	ChatGPT-3.5 Turbo
2	With	OpenAI Embeddings w/ L2	Initial	3	0.5	ChatGPT-3.5 Turbo
3	With	OpenAI Embeddings w/ Cosine	Initial	3	0.5	ChatGPT-3.5 Turbo
4	With	Hybrid Search w/ L2	Initial	3	0.5	ChatGPT-3.5 Turbo
5	With	Hybrid Search w/ Cosine	Initial	3	0.5	ChatGPT-3.5 Turbo
6	With	BM25	Initial	3	0.5	ChatGPT-3.5 Turbo
7	With	Hybrid Search w/ Cosine	First Altered	3	0.5	ChatGPT-3.5 Turbo
8	With	Hybrid Search w/ Cosine	Second Altered	3	0.5	ChatGPT-3.5 Turbo
9	With	Hybrid Search w/ Cosine	Second Altered	5	0.5	ChatGPT-3.5 Turbo
10	With	Hybrid Search w/ Cosine	Second Altered	5	0.25	ChatGPT-3.5 Turbo
11	With	Hybrid Search w/ Cosine	Second Altered	5	0.75	ChatGPT-3.5 Turbo
12	With	OpenAI Embeddings w/ L2	Initial	2	0.5	ChatGPT-4o
13	With	Hybrid Search w/ Cosine	Second Altered	4	0.5	ChatGPT-4o

3.3 Evaluation

We opted for a two-staged approach to the evaluation. For the first stage, two synthetic datasets were created: a close-ended dataset comprising 251 question-answer pairs and a dataset with 50 open-ended questions. The first dataset was generated from documents in the knowledge base to create questions directly related to the content (e.g. restaurants in Lisbon), ensuring the chatbot could effectively retrieve and understand the information provided. The second dataset focused on common tourist questions requesting recommendations, which is crucial for a touristic use case. Note that a human evaluation method would not be feasible at this stage due to the significant time and cost required for comprehensive human evaluations for each experiment. These experiments were conducted on an Acer Swift SFX16-51G with 16 GB of RAM and an Intel Core i7-11390 H processor. Each experiment was repeated 30 times to account for randomness in the responses and time measurements. Time measurements were conducted exclusively for the Chroma experiment, as its primary focus was on reducing response time.

For the second stage, three experts, two from Turismo de Portugal and one from a private tourism company, tested the best-performing version of the chatbot from the experiments. The evaluation focused on three key areas: information, recommendation and engagement. This allowed the chatbot to be tested in a way that closely

resembles its downstream task, with experts directly simulating tourist inquiries and evaluating the chatbot based on their experience.

3.3.1 Automatic evaluation

3.3.1.1 Close-ended The close-ended dataset was generated using LlamaIndex Question Generation tools (LlamaIndex, 2024). A synthetic dataset was created by generating questions for each set of documents, which were then manually verified and provided with corresponding answers. Additional questions were also created to ensure all types of documents were covered. The intention behind this dataset was to include highly specific questions with answers that are not subject to interpretation. The questions vary in format and can be as simple as yes/no inquiries or as detailed as descriptive questions. Generally, the tool can generate any type of question that can be answered by the data in the document. This is very valuable for measuring the performance of the RAG workflow, as it directly addresses knowledge available for retrieval. This approach is far more efficient than manually reviewing each document and creating a diverse set of questions.

For example, consider a document from the Visit Lisboa website about the Macau Scientific and Cultural Centre, which states: "It also strives to highlight the importance of the thousand-year-old heritage of the Chinese civilization and it is one of the most important collections of Chinese art in the Iberian Peninsula." The question generated for this document was: "Which museum has one of the most important collections of Chinese art?" with the answer being "Macau Scientific and Cultural Centre". This dataset was also divided in categories (e.g. Metro) and sub-categories (e.g. Art), based on the origin of the data, to help identify where the problems could be and guide the additional preprocessing.

The system prompt changed for this evaluation, but still followed the foundational principles of Prompt Engineering (Chen et al. 2023): role-prompting (question answering chatbot for touristic information), providing clear and precise instructions (how it should answer and what it will have access to) and having an example. The remainder of the RAG workflow was kept unchanged to effectively test the chatbot's performance. The system prompt adjustment was made to ensure the chatbot's behavior remained consistent with the evaluation approach by changing the instructions given to emphasize the question answering goal and the need for correct answers. The answers in this dataset were evaluated based on Semantic Answer Similarity (SAS), utilizing a cross-encoder architecture with the language model cross-encoder/stsb-roberta-large. SAS evaluates how semantically similar two answers are by analyzing their meaning rather than relying solely on word-for-word comparisons. Technically, SAS works by using a cross-encoder architecture instead of calculating separate embeddings for the input texts (as in the bi-encoder architecture). In the cross-encoder approach, the two texts are concatenated with a special separator token in between them. This metric outperformed lexical-based (e.g. BLEU) and semantic-based (e.g. BERTScore trained) similarity of answer pairs (Risch et al. 2021).

3.3.1.2 Open-ended The open-ended dataset was generated with the help of the ChatGPT-3.5 Turbo model (again, due to the recent release of ChatGPT-4o), with the user prompt:

> What questions would a tourist ask an AI assistant designed to provide information to tourists in Lisbon?

Only this prompt was given to the model, without any system prompt or integration into the RAG workflow previously mentioned. These questions were, again, verified manually and some were added to ensure the model could make recommendations for the most common scenarios in Lisbon. Using ChatGPT-3.5 Turbo allows for the rapid generation of diverse questions based on its extensive training data, ensuring coverage of the typical touristic scenarios, while avoiding the time and resource constraints of collecting these queries from potential users. The intention behind this dataset was to have open-ended questions, where the model did not have an expected answer. As an example of a question in this dataset: Are there any famous markets in Lisbon?

The system prompt used for this evaluation approach is identical to the one designed for the chatbot model, with the only difference being that if the answer cannot be found in the provided information, it should respond with: "I am sorry but I could not find an answer! Is there any other way I can help you?" This ensures that the model acknowledges when it cannot locate the requested information in the documents. The evaluation metric for this dataset was simply based on whether the model could find an answer (0 if it could not and 1 if it could). With it, a percentage can be calculated. As previously, the remainder of the RAG workflow was not changed.

3.3.2 Human evaluation

After identifying the best-performing version of the chatbot, we deployed it locally using the Streamlit library (Streamlit, 2024) and then invited 3 experts from Turismo de Portugal to test the chatbot. A Google Form was created with four sections: Recommendation, Information, Engagement and Overall Experience. As mentioned earlier, the goal of the chatbot is to recommend places to visit, tourist routes and local events; provide general information about Lisbon and engage meaningfully with visitors. Consequently, the evaluation was aligned with these objectives.

Currently, there are no standardized methods or universally agreed-upon best practices for evaluating LLM-based systems (Abeysinghe and Circi 2024). However, we believe that the evaluation process should closely reflect the intended downstream tasks. Thus, the sections and questions chosen were determined based on the objective of the chatbot. Table 3 shows the questions present in the survey, along with the possible responses.

The Recommendation section assessed the chatbot's ability to provide relevant and helpful suggestions tailored to user preferences and needs. The Information section evaluated the accuracy, clarity and reliability of the information provided. The Engagement section focused on user interaction, including the chatbot's conversa-

Table 3 Survey questions and response options

Section	Question	Type	Options
Recommendation	Did Lisa provide helpful recommendations for places to visit in Lisbon?	Multiple choice	Yes or No
	Was Lisa able to make recommendations based on your preferences or needs?	Multiple choice	Yes or No
	Was Lisa able to adapt its recommendations based on a specific time period?	Multiple choice	Yes or No
	How relevant were the recommendations given by the Lisa to your interests?	Likert scale with options for activities, restaurants and hotels	1 (Very Irrelevant) to 5 (Very Relevant) or Not Tested
	How likely are you to follow Lisa's recommendation?	Likert scale	1 (Very Unlikely) to 5 (Very Likely)
	Any additional comments?	Open-ended	N/A
Information	Was the information provided by Lisa accurate and up-to-date (until December 2023)?	Multiple choice	Yes or No
	Did Lisa answer your questions clearly and completely?	Multiple choice	Yes or No
	How confident are you that the information provided by Lisa is reliable?	Likert scale	1 (Not Confident) to 5 (Very Confident)
	Any additional comments?	Open-ended	N/A
Engagement	Did Lisa maintain your interest throughout the conversation?	Multiple choice	Yes or No
	Did you feel Lisa's personality (e.g., tone, friendliness) was appealing and appropriate?	Multiple choice	Yes or No
	Was Lisa able to maintain a coherent conversation on the same topic for an extended period?	Multiple choice	Yes or No
	How quickly did the chatbot respond to your inquiries?	Likert scale	1 (Very Slowly) to 5 (Very Quickly)
	How would you rate the overall user experience when chatting with Lisa?	Likert scale	1 (Very Dissatisfied) to 5 (Very Satisfied)
	Any additional comments?	Open-ended	N/A
Overall Experience	How do you rate your overall satisfaction with Lisa?	Likert scale	1 (Very Dissatisfied) to 5 (Very Satisfied)
	Do you believe Lisa could be helpful for tourists?	Open-ended	N/A

tional coherence, tone and responsiveness. Lastly, the Overall Experience section captured users' general satisfaction and perceptions of the chatbot's utility.

Questions were formatted as multiple-choice, Likert scale or open-ended to capture a mix of quantitative and qualitative feedback. Hence, the survey should ensure a comprehensive assessment of the chatbot's performance across its primary use cases.

Table 4 Performance with and without chroma

Chroma	Close-ended		Open-ended	
	SAS (%)	Time	Score (%)	Time
Without	55.75	7.39	93.80	9.04
With	50.72	0.75	93.08	3.06

Table 5 Performance with the different retrieval algorithms

Algorithm	Close-ended	Open-ended
	SAS (%)	Score (%)
OpenAI Embeddings w/ L2	50.72	93.08
OpenAI Embeddings w/ Cosine	52.97	92.60
Hybrid Search w/ L2	54.75	92.16
Hybrid Search w/ Cosine	56.66	94.20
BM25	49.87	90.60

Table 6 Performance with the different datasets

Dataset	Close-ended	Open-ended
	SAS (%)	Score (%)
Initial	56.66	94.20
First Altered	59.87	94.46
Second Altered	65.73	94.20

4 Results

4.1 Automatic evaluation

Initially, the impact of a vector database (Chroma) was examined, considering it should reduce the time needed for the document retrieval significantly. The algorithm utilized was the one based on OpenAI Embeddings, chosen for its simplicity. The metrics and time obtained are presented in Table 4.

The results revealed a trade-off in accuracy, particularly in the close-ended evaluation (around 5% points—from 55.75 to 50.72%), and the significant time advantage offered by Chroma, which was around three times faster for open-ended queries and around 10 times faster for close-ended. Considering the very strong time difference, Chroma was used for all subsequent tests. Afterwards, the various retrieval algorithms were tested, given that they should be one of the most important factors affecting the performance of the model. The results are described in Table 5.

Results show that Hybrid Search with cosine similarity achieved the highest performance—56.66% SAS and a 94.20% open-ended score—while BM25 was the lowest at 49.87% and 90.60%, respectively. Therefore, Hybrid Search with cosine similarity will be the algorithm used for all the next experiments. As a next step, the different preprocessed datasets were tested to check whether the simplification and remotion of unnecessary content contributed to gains in performance. Table 6 shows the metrics obtained.

There was a significant change in SAS (around 9% points—from 56.66 to 65.73%) while the open-ended score is very similar, always around 94%. As a result, the most preprocessed dataset will be utilized. Then, as the last dataset has less tokens per

Table 7 Performance comparison based on the number of documents retrieved

Documents	Close-ended	Open-ended
	SAS (%)	Score (%)
3	65.73	94.20
5	67.34	94.53

Table 8 Performance with different temperature parameters

Temperature	Close-ended	Open-ended
	SAS (%)	Score (%)
0.25	67.23	94.20
0.5	67.34	94.53
0.75	66.85	95.00

Table 9 Performance comparison based on the two chatbot models

Case		Close-ended	Open-ended
		SAS (%)	Score (%)
Baseline	ChatGPT-3.5 Turbo	50.72	93.08
Baseline	ChatGPT-4o	57.23	92.40
Final Model	ChatGPT-3.5 Turbo	67.34	94.53
Final Model	ChatGPT-4o	71.36	95.20

document, the number of retrieved documents can be increased from 3 to 5 without surpassing the context limit. The close-ended and open-ended evaluations are represented in Table 7.

There were slight improvements in both close-ended (from 65.73 to 67.34%) and open-ended questions (from 94.20 to 94.53%), thus 5 documents will now be used. Finally, the temperature parameter was evaluated to see if there were relevant changes in the metrics as can be seen in Table 8.

Across the tested temperature settings, the close-ended evaluation averaged around 67%, while the open-ended evaluation varied between 94% and 95%. As no significant variations were observed between the metrics, we selected 0.5 as the final temperature parameter. This value represents a midpoint that balances accurate question answering with the generation of creative responses.

In summary, the best performing experiment for ChatGPT-3.5 Turbo utilized the following parameters: Chroma as a vector database, Hybrid Search with Cosine as the retrieval algorithm, the second altered dataset, 5 documents and a 0.5 temperature value. The final two experiments compared ChatGPT-3.5 Turbo and ChatGPT-4o for the baseline case and the case with the best results for the initial model, as described in the Sect. 3.2.4. The results obtained are presented in Table 9.

For the baseline, the close-ended evaluation showed an improvement, increasing from 50.72 to 57.23%, while the open-ended evaluation produced similar scores, around 92–93%. For the best scenario, there was again an improvement in the close-ended evaluation, increasing from 67.34 to 71.36%, while the open-ended remained consistent around 95%. The previous experiments led to an overall improvement in the ChatGPT-4o model's performance by 14% points for the close-ended evaluation (from 57.23 to 71.36%) and 3% points for the open-ended evaluation (from 92.40 to 95.20%).

4.2 Human evaluation

After completing the automatic evaluation, the best-performing conversational system was identified: ChatGPT-4o, combined with Chroma, Hybrid Search utilizing Cosine as the retrieval algorithm, the second altered dataset as the knowledge base, four documents and a temperature of 0.5. This configuration was then deployed locally for expert testing. Figure 3 illustrates an example of a conversation with the developed chatbot.

The results per section of the survey will now be presented, with Portuguese answers translated to English. The Recommendation section results are described in Table 10.

All experts agreed that Lisa provided helpful recommendations tailored to their needs and specified time periods. Recommendation relevance across all categories (activities, restaurants and hotels) ranged from "Relevant" to "Very Relevant." Experts reported they were "Likely" to "Very Likely" to follow Lisa's suggestions. Comments highlighted its comprehensive and natural responses, providing suggestions for further enhancements, which will be further discussed in the following section. Table 11 shows the results for the Information section.

The chatbot was rated as accurate, up-to-date and clear in its responses, with one expert noting the reliability of its information based on their tourism knowledge. The experts rated the reliability of the answers between "Confident" and "Very Confident". Table 12 presents the results for the Engagement section.

Lisa maintained user interest, displayed an appropriate personality and provided coherent and timely responses, resulting in a highly rated overall user experience (all were "Very Satisfied"). The speed of the responses ranged from "Quickly" to "Very Quickly". Lastly, Table 13 outlines the responses given in the Overall Experience section.

All experts expressed high satisfaction with Lisa ("Very Satisfied"), affirming its potential usefulness for tourists.

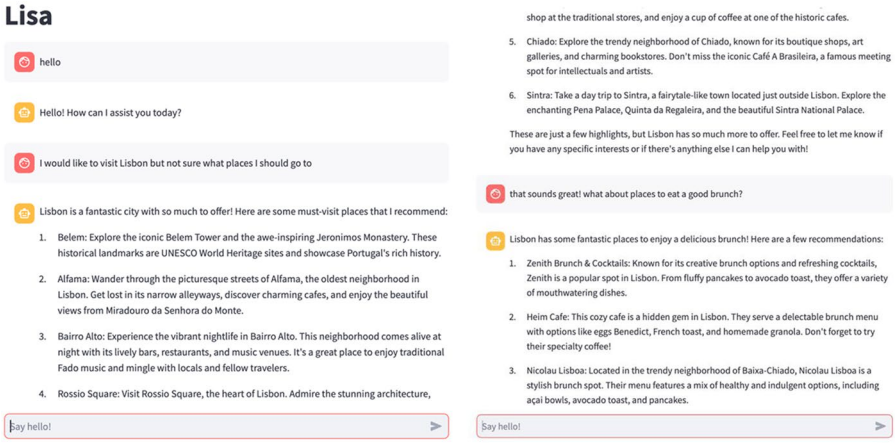

Fig. 3 Example conversation

Table 10 Results for the recommendation section

Question	Expert 1	Expert 2	Expert 3
Did Lisa provide helpful recommendations for places to visit in Lisbon?	Yes	Yes	Yes
Was Lisa able to make recommendations based on your preferences or needs?	Yes	Yes	Yes
Was Lisa able to adapt its recommendations based on a specific time period?	Yes	Yes	Yes
How relevant were the recommendations given by the Lisa to your interests?	Activities: 4 (Relevant) Restaurants: 4 (Relevant) Hotels: 4 (Relevant)	Activities: 4 (Relevant) Restaurants: 5 (Very Relevant) Hotels: 5 (Very Relevant)	Activities: 4 (Relevant) Restaurants: 5 (Very Relevant) Hotels: Not Tested
How likely are you to follow Lisa's recommendation?	4 (Likely)	5 (Very Likely)	5 (Very Likely)
Any additional comments?	Lisa has a highly comprehensive response capability, providing accurate information on historical context, tourist itineraries, tourist activities and even incorporating a touch of emotional intelligence.	Info about payments with debit cards (ex. metro), strikes, suggestions for tourism circuits based on atmospheric conditions, gamification, escape rooms added to circuits, buy tickets directly from the chatbot, offline chatbot. just some more suggestions:)) WELLDONE!!	Works very well! Lisa gives you first a wide answer so you can see "the big picture" and take it from there narrowing down based on more specific questions. Accurate responses and very complete.

Table 11 Results for the information section

Question	Expert 1	Expert 2	Expert 3
Was the information provided by Lisa accurate and up-to-date (until December 2023)?	Yes	Yes	Yes
Did Lisa answer your questions clearly and completely?	Yes	Yes	Yes
How confident are you that the information provided by Lisa is reliable?	4 (Confident)	5 (Very Confident)	4 (Confident)
Any additional comments?	I gave a 4 in this initial phase of Lisa because I didn't have enough time to actually double check and proof the accuracy of her (its?) suggestion. But based on my knowledge in tourism is seems pretty reliable.		Congrats!:)

Table 12 Results for the engagement section

Question	Expert 1	Expert 2	Expert 3
Did Lisa maintain your interest throughout the conversation?	Yes	Yes	Yes
Did you feel Lisa's personality (e.g., tone, friendliness) was appealing and appropriate?	Yes	Yes	Yes
Was Lisa able to maintain a coherent conversation on the same topic for an extended period?	Yes	Yes	Yes
How quickly did the chatbot respond to your inquiries?	4 (Quickly)	5 (Very Quickly)	5 (Very Quickly)
How would you rate the overall user experience when chatting with Lisa?	5 (Very Satisfied)	5 (Very Satisfied)	5 (Very Satisfied)
Any additional comments?			Keep improving:)

Table 13 Results for the overall experience section

Question	Expert 1	Expert 2	Expert 3
How do you rate your overall satisfaction with Lisa?	5 (Very Satisfied)	5 (Very Satisfied)	5 (Very Satisfied)
Do you believe Lisa could be helpful for tourists?	Lisa seems to be a great tool that will assist tourists accurately.	SURE !	

5 Discussion

The initial evaluation of Chroma demonstrated a significant reduction in document retrieval time, aligning with the efficiency and trade-offs noted in the literature (Han et al. 2023). While a decrease in accuracy was observed, the substantial improvement in speed justified its adoption in all subsequent tests. This highlights the practical advantage of using Chroma for chatbot use cases, where response time is crucial.

The experiments with different retrieval algorithms revealed that Hybrid Search, particularly when used in conjunction with cosine similarity, offered the best performance. Hybrid Search allowed for more robust and accurate results because it combines the strengths of both sparse and dense retrieval methods (LangChain, 2024). Additionally, the superior performance of cosine similarity over L2 norm aligns with relevant literature, which identifies cosine similarity as one of the best measures for text retrieval (Thompson et al. 2015).

Preprocessing steps proved to be highly beneficial, resulting in a notable improvement in performance. This improvement surpasses the gains achieved by changing the retrieval algorithm, reflecting the importance of data quality. A relevant example from the literature is the development of phi-1, a small language model that competes effectively with much larger models by leveraging a high-quality dataset (Gunasekar et al. 2023).

Increasing the number of retrieved documents from 3 to 5 yielded slight improvements in performance. As the answers to the questions are present in only one document, multiple documents are not necessary to arrive at the correct answer, justifying the small difference.

The evaluation of the temperature parameter revealed no significant variations. This minimal effect suggests that the influence of temperature is negligible despite the potential for increased randomness in outputs. This is likely due to the additional context, which leaves little room for errors in the response.

The comparison between ChatGPT-3.5 Turbo and ChatGPT-4o highlighted advancements in model capabilities, with ChatGPT-4o showing strong improvements. These results suggest that ChatGPT-4o is more effective at interpreting context. In the final model evaluation, ChatGPT-4o showed an improvement in 14% points for close-ended tasks and in 3% points for open-ended tasks compared to ChatGPT-3.5 Turbo. This suggests that even current ChatGPT models need optimization for specific downstream tasks, aligning with the findings on optimization. Another important observation is the increased consistency in outputs of the ChatGPT-4o model compared to ChatGPT-3.5 Turbo, with fewer instances of hallucination (though this was verified manually). This supports the findings from the GPT-4 Technical Report from OpenAI et al. (2023).

The survey results demonstrate that Lisa performed consistently well across all evaluated sections, effectively meeting its recommendation, information and engagement goals. The chatbot showcased its ability to assist tourists by processing complex information, facilitating natural interactions and generating appropriate recommendations, as highlighted by the experts, surpassing the current limitations of existing touristic chatbot implementations. These findings further validate the robustness of this generative methodology.

Additionally, one expert provided valuable suggestions for improvement. These included addressing specific information needs, such as strikes, which can significantly impact tourists relying on public transportation and who may be unaware of local disruptions. Information on payments with debit cards for public transportation was also highlighted as critical, particularly in cases where tourists may need to acquire local currency. The suggestion to adapt touristic circuits based on atmospheric conditions is particularly important, as weather strongly influences recommendations. Technically, this could be implemented by integrating a function to retrieve and process external weather data.

Other suggestions focused on product development beyond conversational functionality, including gamification to enhance user engagement, offline functionality to accommodate tourists with limited mobile data and the ability to purchase tickets directly through the chatbot. These insights provide actionable directions to improve Lisa's overall capabilities.

6 Conclusion

6.1 Contributions

In this research, we developed a general-purpose chatbot designed to address the unique needs of tourists visiting Lisbon. The chatbot recommends places to visit, tourist routes and local events, provides general information about the city and engages with users through natural interactions. To achieve this, we constructed a compre-

hensive knowledge base by web-scraping over 2000 website pages and designed two synthetic evaluation datasets to iteratively improve the model's performance. The chatbot was evaluated using both automatic and human-based methods, achieving strong results across recommendation, information and engagement use cases.

In general, the results clearly indicated the need for optimizing ChatGPT models for specific use cases and adopting an iterative approach similar to other data science problems. Optimization can come from various areas, such as enhancing the data and refining the RAG pipeline. One of the most crucial factors is the quality of the knowledge base, as evidenced by the score variations obtained, which is expected since the model relies on the knowledge base to produce accurate answers.

However, there are trade-offs to consider in optimization when moving into production, such as the cost of these models and latency issues, as illustrated in the use of Chroma. As noted by Barnett et al. (2024), these RAG systems need to be tested and monitored over time in production. The optimization process needs to be a continuous effort.

Additionally, the evaluation for the experiments was driven by synthetic data, demonstrating how it can quickly generate test sets tailored to specific use cases, which is essential for downstream tasks. This involves mapping out how the model will be used, allowing for the creation of relevant synthetic data. In this case, the two areas tested were the model's recommendation and question-answering capabilities.

The two-stage evaluation approach demonstrates how synthetic and human evaluations can complement each other. Synthetic data provides a controlled environment for iterative experimentation, while human feedback offers actionable insights, strongly aligned with the downstream task, that can inform further refinements. Combining these methods provides a more comprehensive assessment of the chatbot's overall performance.

This study makes several key contributions to the development of generative touristic chatbots for cities.

First, by implementing a large language model based architecture, we directly address critical limitations that have historically plagued tourism chatbots, specifically their inability to handle complex and contextual dialogue and generate truly personalized responses rather than generic information.

Second, our systematic optimization methodology demonstrates how each system component can be iteratively refined and measured, providing a scalable framework that moves beyond ad-hoc development approaches.

Third, our dual-phase evaluation methodology combines rapid synthetic data testing for system components with human assessment that captures nuanced feedback aligned with real application scenarios. These methodological advances provide a structured approach for developing tailored conversational systems that can be replicated across different tourism contexts and related domains.

Fourth, we show how an advanced generative chatbot – such as our prototype Lisa – can be a significant innovation within the tourism sector by seamlessly integrating tourists with local services through scalable and adaptive solutions. Functioning as a virtual concierge available 24/7, the chatbot has the capacity to alleviate the operational burdens faced by traditional tourism offices while enabling destinations to manage increasing visitor volumes with greater efficiency. Through the provi-

sion of localized information - spanning transportation schedules, event updates and emergency alerts - the chatbot ensures that tourists are equipped with the necessary knowledge to make informed and confident decisions throughout their journeys. Unlike earlier studies that focus primarily on rule-based or narrow-domain chatbots, our methodology leverages a large, web-scraped knowledge base and optimized prompts to deliver nuanced, context-aware guidance - providing general information, handling complex queries and offering a stronger conversational experience, as demonstrated by the evaluation results. Furthermore, its multilingual capabilities and sensitivity to cultural nuances enhance accessibility, promoting inclusivity and broadening the global appeal of destinations such as Lisbon to diverse audiences. In addition to facilitating visitor engagement, the chatbot also supports sustainable tourism management by promoting environmentally conscious practices. For instance, it can recommend eco-friendly travel options such as public transportation, pedestrian routes, and bicycle rentals, thereby fostering sustainable mobility. Moreover, the chatbot's capacity to educate tourists on local sustainability initiatives, cultural etiquette, and environmental conservation measures plays a critical role in enhancing visitors' awareness and accountability. These capabilities underscore the potential of chatbots like Lisa not only to enrich the tourist experience but also to contribute meaningfully to the broader objectives of sustainable destination management.

6.2 Limitations and future work

Despite these contributions, this work has some limitations that need to be considered and discussed.

One limitation of this work is that the chatbot was not tested by actual tourists in Lisbon. Before deploying the system in a production environment, it would be essential to involve real users, ideally with a large and diverse sample, to evaluate its performance. Additionally, the inclusion of separate validation and test sets and the participation of human annotators in the development of the evaluation datasets could improve the generalizability of the scores.

Another limitation is that ChatGPT-4o was tested in only two scenarios and improvements made for ChatGPT-3.5 Turbo may not directly apply to ChatGPT-4o. However, the increased scores in ChatGPT-4o suggest that most decisions can be transferred effectively.

For future work, an expanded exploration of optimization would be beneficial, including different models for both retrieval and text generation, better structures for the web-scraped data and other ways to improve the RAG pipeline, considering both cost and time. It would also be interesting to explore employing large language models as autonomous agents, significantly enhancing the chatbot's functionality (Xi et al. 2023) - for example, by enabling tasks such as making reservations or purchasing tickets, as suggested by the experts. Integrating an expanded knowledge base that includes real-time information on transit strikes, weather forecasts and alerts and accepted payment methods for public transportation would further improve its utility. By embedding this agent-enabled chatbot within an official government or municipal platform and establishing partnerships with external service providers, users would

be able to plan, reserve and complete every phase of their tourist journey in a single, seamless interaction.

Furthermore, exploring how these systems can adapt to different users and consider accessibility would be important, as Cena et al. (2023) highlights. This could be addressed through an improved knowledge base with details around accessibility and different types of users or better prompting that considers from the start the type of user and how to best work around their needs.

Specifically, although tested manually by the authors, investigating how the inclusion of different languages within the knowledge base impacts retrieval quality would be valuable, as there seemed to be a preference for English retrieval when the query was also in English. Also, the model's comparatively lower performance in Portuguese further emphasizes the need to test other models better suited for this specific language. Finally, despite the added sentence specifying the current time in the system prompt, exploring how the model could better interpret temporal prompts to enhance its ability to make time-based recommendations, such as events occurring within the current month, would be worthwhile.

Author contributions M.C. developed the models and M.C. and B.J. wrote the main manuscript text. All authors reviewed the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on). This work was funded by Portuguese national funds through the Portuguese Foundation for Science and Technology—FCT under research grant 2024.07501.IACDC-Inteligência Artificial, Ciência dos Dados e Cibersegurança de relevância na Administração Pública and research grant FCT UIDB/04152/2020–Centro de Investigação em Gestão de Informação (MagIC). This work is also funded by PRR – Plano de Recuperação e Resiliência and by the NextGenerationEU funds, through the scope of the Agenda for Business Innovation "ATT – Agenda Mobilizadora Acelerar e Transformar o Turismo" (Project no. 47 with the application C645192610-00000060).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Abeysinghe B, Circi R (2024) The challenges of evaluating LLM applications: an analysis of automated, human, and LLM-based approaches. http://arxiv.org/abs/2406.03339

Alhumoud S, Diab A, Aldukhai D, Alshalhoub A, Alabdullatif R, Alqahtany D, Alalyani M, Bin-Aqeel F (2022) Rahhal: a tourist Arabic Chatbot. Proceedings –2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 66–73. https://doi.org/10.1109/SMARTTECH54121.2022.00028

Amalia A, Suprayogi MS (2019) Engaging millennials on using chatbot messenger for eco-tourism. Proc Third Int Conf Sustainable Innov 2019 – Humanity Educ Social Sci (IcoSIHESS 2019) 484–487. https://doi.org/10.2991/icosihess-19.2019.84

Anand S, Abhishek Sai AM, Karthikeya M (2023) Chatbot Enabled Smart Tourism Service for Indian Cities: an AI Approach. 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, IEMECON 2023. https://doi.org/10.1109/IEMECON56962.2023.10092286

Arteaga D, Arenas J, Paz F, Tupia M, Bruzza M (2019) Design of information system architecture for the recommendation of tourist sites in the city of Manta, Ecuador through a Chatbot. 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 1–6. https://doi.org/10.23919/CISTI.2019.8760669

Barnett S, Kurniawan S, Thudumu S, Brannelly Z, Abdelrazek M (2024) Seven failure points when engineering a retrieval augmented generation system. 2024 IEEE/ACM 3rd Int Conf AI Eng – Softw Eng AI (CAIN) 2022–March:194–199. https://doi.org/10.1145/3644815.3644945

Boné J, Ferreira JC, Ribeiro R, Cadete G (2020) Disbot: a Portuguese disaster support dynamic knowledge chatbot. Appl Sci 10(24):1–20. https://doi.org/10.3390/app10249082

Bulchand-Gidumal J, Secin W, O'Connor E, P., Buhalis D (2024) Artificial intelligence's impact on hospitality and tourism marketing: exploring key themes and addressing challenges. Curr Issues Tourism 27(14):2345–2362. https://doi.org/10.1080/13683500.2023.2229480

Caldarini G, Jaf S, McGarry K (2022) A literature survey of recent advances in chatbots. Information 13(1). https://doi.org/10.3390/info13010041

Calvaresi D, Ibrahim A, Calbimonte J, Schegg R, Fragnière E, Schumacher MI (2021) The evolution of Chatbots in tourism: a systematic literature review. Information and Communication Technologies in Tourism 2021

Casillo M, Clarizia F, D'Aniello G, De Santo M, Lombardi M, Santaniello D (2020) CHAT-Bot: a cultural heritage aware teller-bot for supporting touristic experiences. Pattern Recognit Lett 131:234–243. https://doi.org/10.1016/j.patrec.2020.01.003

Cena F, Mauro N, Ardissono L, Ferrero F, Ferrigno S, Rapp A, Mattutino C, Keller R (2023) CARES: an inclusive personalized touristic system for autism. UMAP 2023 - Adjun Proc 31st ACM Conf User Model Adaptation Personalization 363:366. https://doi.org/10.1145/3563359.3596665

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X (2023) A survey on evaluation of large Language models. ACM Trans Intell Syst Technol 15:1–45. http://arxiv.org/abs/2307.03109

Chen B, Zhang Z, Langrené N, Zhu S (2023) Unleashing the potential of prompt engineering in large language models: a comprehensive review. http://arxiv.org/abs/2310.14735

Chroma. (n.d.). Chroma. Retrieved February 1 (2024) from https://www.trychroma.com/

Cormack GV, Clarke CL, Büttcher S (2009) Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval

Doğruöz AS, Skantze G (2022) How open are the conversations with open-domain chatbots? A proposal for speech event based evaluation. http://arxiv.org/abs/2211.13560

Fraiwan M, Khasawneh N (2023) A review of ChatGPT applications in education, marketing, software engineering, and healthcare. Benefits, Drawbacks, and Research Directions

Gunasekar S, Zhang Y, Aneja J, Mendes CC, Giorno AD, Gopi S, Javaheripi M, Kauffmann P, de Rosa G, Saarikivi O, Salim A, Shah S, Behl HS, Wang X, Bubeck S, Eldan R, Kalai AT, Lee YT, Li Y (2023) Textbooks are all you need.

Han Y, Liu C, Wang P (2023) A comprehensive survey on vector database: storage and retrieval technique, Challenge. http://arxiv.org/abs/2310.11703

Kenney NM (2023) A brief analysis of the architecture, limitations, and impacts of ChatGPT. https://doi.org/10.5281/zenodo.7762245

Kusal S, Patil S, Choudrie J, Kotecha K, Mishra S, Abraham A (2022) AI-based conversational agents: A scoping review from technologies to future directions. IEEE Access 10:92337–92356. https://doi.org/10.1109/ACCESS.2022.3201144

LangChain. (n.d.). Ensemble retriever. Retrieved February 1 (2024) from https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/ensemble/
Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. http://arxiv.org/abs/2005.11401
Li X, Zhang Y, Malthouse EC (2023) Exploring fine-tuning ChatGPT for news recommendation. http://arxiv.org/abs/2311.05850
Lin CC, Huang AYQ, Yang SJH (2023) A review of AI-Driven conversational chatbots implementation methodologies and challenges (1999–2022). Sustainability 15(5). https://doi.org/10.3390/su15054012
LlamaIndex. (n.d.). QuestionGeneration. Retrieved February 1 (2024) from https://docs.llamaindex.ai/en/stable/examples/evaluation/QuestionGeneration/
Lund BD, Wang T (2023) Chatting about chatgpt: how May AI and GPT impact academia and libraries? https://ssrn.com/abstract=4333415
Mijwil M, Hiran M, Doshi KK, Dadhich R, Al-Mistarehi M, A.-H., Bala I (2023) ChatGPT and the future of academic integrity in the artificial intelligence era: a new frontier. Al-Salam J Eng Technol 2(2):116–127. https://doi.org/10.55145/ajest.2023.02.02.015
Mnasri M (2019) Recent advances in conversational NLP: towards the standardization of Chatbot building. http://arxiv.org/abs/1903.09025
Muennighoff N, Tazi N, Magne L, Reimers N (2022), October 13 MTEB: massive text embedding benchmark. Conference of the European Chapter of the Association for Computational Linguistics. http://arxiv.org/abs/2210.07316
Nhut Lam K, Huu Nguy L, Lam VLE, Kalita J (2023) A Transformer-based educational virtual assistant using diacriticized Latin script. IEEE Access 10:92337–92356. https://doi.org/10.1109/ACCESS.2023.3307635
OpenAI (2023), November 13 A survey of techniques for maximizing LLM performance [Video recording]. https://www.youtube.com/watch?v=ahnGLM-RC1Y
OpenAI. (n.d.). API Documentation. Retrieved February 1 (2024) from https://platform.openai.com/docs/introduction
OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Zoph B (2023) GPT-4 Technical Report. http://arxiv.org/abs/2303.08774
OpenAI,:, Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow A, Welihinda A, Hayes A, Radford A, Mądry A, Baker-Whitcomb A, Beutel A, Borzunov A, Carney A, Chow A, Kirillov A, Malkov Y (2024) GPT-4o System Card. http://arxiv.org/abs/2410.21276
Risch J, Möller T, Gutsch J, Pietsch M (2021) Semantic answer similarity for evaluating question answering models. http://arxiv.org/abs/2108.06130
Selenium. (n.d.). WebDriver Documentation. Retrieved February 1 (2024) from https://www.selenium.dev/documentation/webdriver/
Silva B, Nunes L, Estevão R, Aski V, Chandra R (2023) GPT-4 as an Agronomist Assistant? answering agriculture exams using large language models. http://arxiv.org/abs/2310.06225
Sperlí G (2021) A cultural heritage framework using a deep learning based chatbot for supporting tourist journey. Expert Syst Appl 183. https://doi.org/10.1016/j.eswa.2021.115277
Streamlit. (n.d.). Documentation. Retrieved April 15 (2024) from https://docs.streamlit.io/
Svore KM, Burges CJC (2009) A machine learning approach for improved BM25 retrieval. Int Conf Inform Knowl Manage Proc 1811–1814. https://doi.org/10.1145/1645953.1646237
Thompson VU, Panchev C, Oakes MP (2015) Performance evaluation of similarity measures on similar and dissimilar text retrieval. 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 577–584
TravelBI by Turismo de Portugal (2024) Turismo em Números. https://travelbi.turismodeportugal.pt/turismo-em-portugal/turismo-numeros-2023/
Trotman A, Puurula A, Burgess B (2014) Improvements to BM25 and language models examined. Proceedings of the 19th Australasian Document Computing Symposium., 27-28-November-2014. https://doi.org/10.1145/2682862.2682863
Turing AM (1950) Mind a Quarterly Review Psychology and Philosophy I-computing Machinery and Intelligence
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. Neural Information Processing Systems. http://arxiv.org/abs/1706.03762

Visit Lisboa. (n.d.). Visit Lisboa. Retrieved January 12 (2024) from https://www.visitlisboa.com/
Weizenbaum J (1966) ELIZA - A computer program for the study of natural Language communication between man and machine. Commun ACM 9(1):36–45
Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao T, Le, Gugger S, Rush AM (2019) HuggingFace's Transformers: State-of-the-art Natural Language Processing. http://arxiv.org/abs/1910.03771
Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E, Zheng R, Fan X, Wang X, Xiong L, Zhou Y, Wang W, Jiang C, Zou Y, Liu X, Gui T (2023) The Rise and Potential of Large Language Model Based Agents: A Survey. http://arxiv.org/abs/2309.07864

Authors and Affiliations

Miguel Cruz[1] · Bruno Jardim[1] · Miguel de Castro Neto[1]

✉ Miguel Cruz
mcruz@novaims.unl.pt

✉ Bruno Jardim
bjardim@novaims.unl.pt

Miguel de Castro Neto
mneto@novaims.unl.pt

[1] NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, Lisboa 1070-312, Portugal