NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

NOVA IMS Assistant

Enhancing Information Access and Campus Engagement through an Intelligent Chatbot

Joana Pardelha Marcelo Sousa

Project Work

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA IMS Assistant

Enhancing Information Access and Campus Engagement through an Intelligent Chatbot

by

Joana Pardelha Marcelo Sousa

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Miguel de Castro Neto, PhD, NOVA Information Management School

Co-Supervised by

Bruno Jardim, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 8th of July of 2024]

ABSTRACT

Chatbots have revolutionized human-technology interactions with their remarkable capabilities in various applications, offering intuitive and efficient communication solutions for diverse environments, including academic contexts. This work focuses on leveraging the advancements of natural language processing models and chatbots to develop a GPT-3.5-based chatbot enhanced with Retrieval-Augmented Generation tailored for the NOVA IMS community. The chatbot was built using LangChain for construction and Chroma for vector storage, enabling the chatbot to provide accurate and contextually relevant responses. Two custom datasets were created to conduct the evaluation of multiple aspects of the chatbot's performance, including similarity measure for the Retriever, chunking strategies, and prompt templates, which included both manual review and RAGAS. Overall, the chatbot performs well, providing accurate and relevant replies within the Nova IMS settings. Despite this, qualitative analysis revealed areas for improvement, such as incomplete answers and irrelevant information.

KEYWORDS

Chatbot; RAG; GPT; Natural Language Processing; Artificial Intelligence

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AIML	Artificial Intelligence Markup Language
A.L.I.C.E.	Artificial Linguistic Internet Computer Entity
BoW	Bag-of-Words
CBOW	Continuous Bag-of-Words
GPT	Generative Pre-Trained Transformer
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
RAG	Retrieval Augmented Generator
RNN	Recursive Neural Networks
TF-IDF	Term Frequency-Inverse Document Frequency

1. INTRODUCTION

Artificial Intelligence (AI) is a field with roots dating several decades back. During this time, the field of AI has undergone remarkable evolution and growth. Encompassing several subfields like machine learning and natural language processing, AI has reshaped how humans engage and interact with technology. The advancements in these subfields have led to better and quicker problem-solving and decision-making, as well as have helped streamline automation processes, making the business processes more efficient and creating new and better ones. In a rapidly changing world, the adaptive nature of AI, its promised efficiency and innovation are all factors that have driven it into various areas of our daily lives. Between virtual assistants, autonomous operation of vehicles, and recommendation systems, like the one recommending us music on a music streaming platform, the applications of AI are as diverse as they are impactful. Among these varied applications, there is one important demonstration of how powerful and capable AI is: the development and integration of chatbots.

As a key area of AI, chatbots use and development has grown across numerous industries, providing an intuitive and user-friendly interface that simulates human speech. They are used in diverse settings, such as customer service (Martins et al., 2022), healthcare appointment scheduling (Dammavalam et al., 2022), and administrative procedures for universities and schools (Amin Kuhail et al., 2023), demonstrating their versatility and usefulness.

Significant milestones have marked the evolution of chatbot technology. Recently, the Generative Pre-trained Transformer (GPT) series, and in particular, the release of the GPT-3 and GPT-3.5 models (Brown et al., 2020) stood out among all these improvements. With their capacity to understand and generate human-like text, GPT series started a new conversational AI era. This technology has shown remarkable capabilities in various applications, from enhancing language understanding to creative text generation. GPT-based chatbots, specifically ChatGPT, have gained traction for their ability to go beyond predefined responses. They are able to generate contextually relevant and coherent replies, offering a more natural and engaging user experience.

As GPT-based chatbots are explored, it becomes evident that their dynamic nature is in line with creating custom solutions for intricate scenarios, such as the development of a user-friendly GPT-based Chatbot for NOVA IMS —the central pursuit in this thesis. Previous research has focused on exploring the use of GPT-based chatbots as educational tools as well as FAQ chatbots for students and clients, this work proposes a modern solution to FAQ sections — specifically, the customization of a GPT-based chatbot that can answer questions about Nova IMS. The focus will be on developing a GPT-based chatbot tailored to the specific needs, context, and requirements of NOVA IMS community. The way this chatbot will be integrated can be seen in Figure 1.

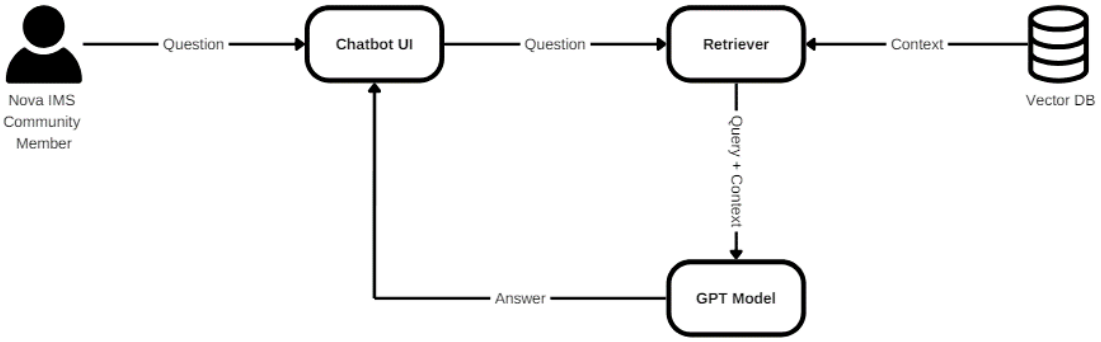

Figure 1 - Integration of a chatbot at Nova IMS

As the main objective of this thesis is the creation of a GPT-based chatbot tailored for the NOVA IMS community, there are secondary aims that will be addressed. These are as follows:

1. Collect and clean documents relating to Nova IMS
2. The creation of datasets to be able to do evaluation for this task.
3. Assess the chatbot's ability to answer questions based on instructions given.

This following work is structured as follows: the Background section will provide the required conceptual and technological context; the Related Work section will explore the use and implementation of chatbots in education and organizations; the Methodology section, that will explain the steps required to create the chatbot; in the Results and Discussion will report and discuss the what was obtained from this work; Conclusions will summarize and conclude what was obtained from this work; Limitations and Recommendations for Future work will the constraints encountered and suggests directions for future research and improvement of the chatbot.

2. BACKGROUND

In this section, the needed background that is the foundation for this work is presented. This chapter aims to provide a comprehensive overview of the key concepts and technologies that make this study possible. The goal is to establish a solid conceptual and technological framework that not only informs the reader but also prepares the groundwork for the project described in the following chapters.

2.1. NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate text in a human-like way (IBM, 2024). It involves the development of machine learning and deep learning models that allow machines to process and analyze and engage in human language in various forms, those being text and speech.

To reach the main goal of understanding and generating language in a human-like form, NLP must surpass some challenges. There are many different languages, which have large vocabularies that can be combined in infinitely many sentences, so the multitude of language and their diversity is a one of the big challenges of NLP. Also, the syntax and semantics differ between languages and within languages themselves, as it happens with regional dialects, which make the scope of what there is to understand much bigger. And to add to that, words have ambiguity as they can have different meanings and often the meaning being used in a sentence is understood through context. Therefore, the semantic relationship between phrases and paragraphs is important to fully understand what is being referenced in text or talked about in speech. This ability feels intuitive to humans but many times even humans get lost in the context or use the wrong syntax and semantics, so to have a computer master this ability is a big challenge (Chowdhary, 2020).

NLP technology has a long history, that can be traced back to when the term “translation machine” was patented in the 1930s (Johri et al., 2021). It started with rule-based systems, which were programs with a list of hard-coded rules that would try to simulate a human conversation. Then, around the 1990s, there was the introduction of machine learning to NLP, which opened door to more complex programs that were able to understand language in a much more comparable way to humans. After that, the use of neural networks was the next big step in NLP. This is the approach that is used by the most popular models and architectures NLP.

One of the most notorious types of neural networks for NLP is the Recurrent Neural Networks (RNN), that were created to process sequential data, like phrases and text (Xiao & Zhou, 2020). They do this by having a hidden state that keeps the information of the previously seen data, making the inputs and outputs dependent on each other, meaning that the previous inputs

can influence the current input and output, crucial when predicting text due to the need to capture context. However, RNNs suffer from a fundamental problem: the vanishing gradient problem. As the learning process progresses and more input is fed to the model, the weights of the data that was seen in the early inputs diminishes. This makes it harder to capture long-term dependencies and leaves a bias towards more recently seen data. To solve this problem the Long Short-Term Memory (LSTM) architecture was proposed by Hochreiter & Schmidhuber (1997). LSTMs tried to solve the problem by introducing a memory cell, that can choose to add or forget information based on the inputs received, which would mean that it would be able to have more important context in the foreground while forgetting less important or useless context. However, this is a very computationally expensive architecture and requires careful fine tuning to achieve satisfactory performance for each task.

The next advancement in NLP was the introduction of Seq2Seq models by Sutskever et al. (2014) with the task of machine learning in mind. In their paper, the architecture is described as a "multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector", which is what is now called the encoder-decoder architecture. The encoder receives the input, processes it, captures the context, and stores it in a hidden state of the network, creating a context vector. This context vector is passed to the decoder which generates the output sequence. At each step, the decoder uses the elements that were already generated, the context vector and the input to make the predictions.

One of the most positive aspects of seq2seq models is that they can handle sentences of different lengths, as in multiple tasks, like machine translation, it is common to not have the input and the output being the same length. Despite all the pros of using seq2seq models, they still have some limitations, like the fact that their context hidden state is of fixed length that can lead to the model having a challenging time capturing long-term dependencies. Like most, seq2seq models require a large amount of training data and are computationally expensive.

To improve the performance of Seq2Seq mechanisms, Bahdanau et al. (2016) introduced attention mechanisms. This mechanism allows the decoder part of an encoder-decoder to generate the output while focusing on distinct parts of the input, prioritizing the most important parts, which improves accuracy and is more computably efficient. Figure 2 shows this process for a translation task in a very simple way.

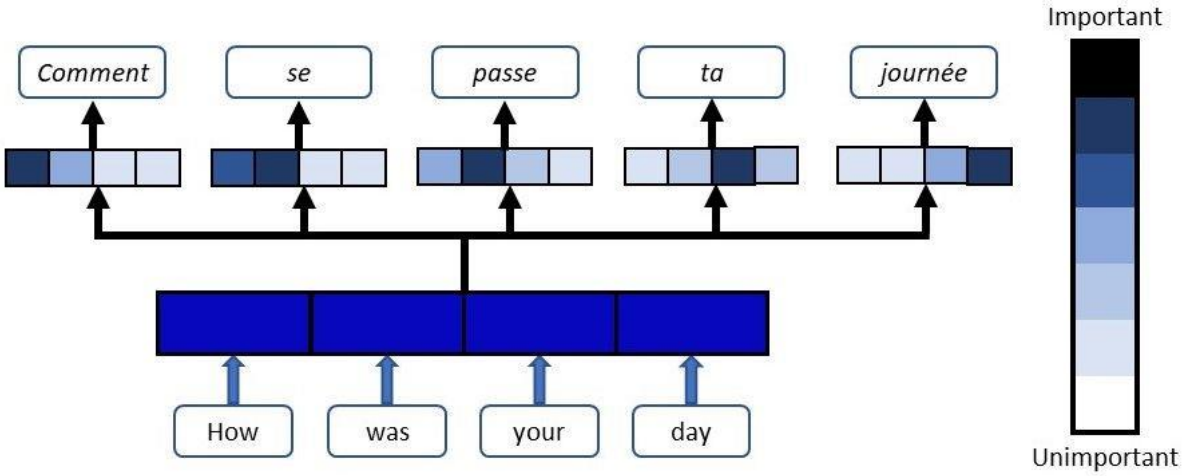

Figure 2 - Simplification of Attention mechanism. Taken from https://blog.floydhub.com/content/images/2019/09/Slide36.JPG

The attention mechanism works by attributing an attention score for each word of the input sequence. It uses three main components: the keys, the values, and the queries, all vectors. It is based on the keys, that the vectors provide a way to represent the relationships between the queries and the elements in the sequence. Then SoftMax is applied to the attention scores obtaining a probability distribution of how much importance to give to each word. These scores are then multiplied by the hidden states of the encoder and the context vector is formed. Like this, the context vector captures the most valuable information from the input sequence. This repeats for each word in the output sequence.

Building on the attention mechanism, another big milestone for NLP was the Transformer architecture by Vaswani et al., (2017). However, this architecture's attention works in a different way. It is named self-attention since each word does have an influence on the attention scores of all other words in the sequence, including the present word itself.

There is also a multi-head attention mechanism, that, as the name hints at, means that we have multiple lines of attention running in parallel, and each will be learning from a different part of the input. Like this, the model captures more relationships in the data, more complex patterns, and more dependencies.

The transformer architecture has three different uses of multi-head attention:

- As encoder-decoder attention layers, where the queries come from the past decoder layer, and the keys and values are from the encoder's output. So, at each decoder position, there is access to all positions of the input sequence.
- As a self-attention layer in the encoder, where all queries, keys, and values come from the same past encoder layer's output, allowing each position in the encoder to attend all other positions within that same past layer.

- As a self-attention layer in the decoder as well. Operating similarly to the encoder self-attention, where at each decoder step, there is access to all the positions from the past layer. This layer is usually masked to ensure that each position can only obtain information from previous positions and not future ones.

The Transformer architecture has a stacked encoder-decoder structure. Each encoder includes the previously mentioned multi-head attention and feedforward neural network. While the decoder has the same two sub-layers as the encoder, it introduces a third sub-layer that conducts the multi-head attention over the encoder's stack output. To not allow the current or future output to have a say in predicting the output, there is a need to partially mask the output.

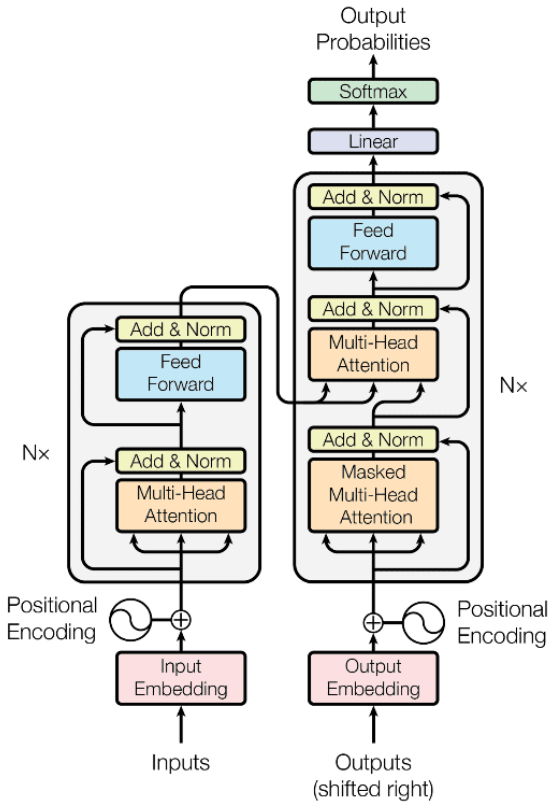

Figure 3 - Transformer architecture. Taken from https://arxiv.org/pdf/1706.03762

Lastly, the last decoder's output goes through a linear neural network and a SoftMax layer. This SoftMax layer transforms the output vector into probabilities for the vocabulary. The full transformer architecture is represented in Figure 3.

This architecture is the basis for many of the most famous state-of-the-art language models. Between those are the GPT series, that will be explained in detail in Section 2.4, the BERT model, and its variants, among many others.

Google developed the BERT model, which stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018). This model is pre-trained with large amounts of data and enables bidirectional movement, which means that can capture context from words that are behind as well as in front of the current word. This improves the understanding and representation of context, syntax, and sematic relationships. BERT has led to incredible breakthroughs in most NLP tasks and is commonly used as a benchmark for other models. From BERT many variants were created such as RoBERTa (Liu et al., 2019), that does the masking during training, or ALBERT (Lan et al., 2020), that is a much smaller model, providing a more efficient computing at low performance price. These models are often fine-tuned into specific tasks to perform at a state-of-the-art level.

There are diverse applications of NLP, some are even part of the day to day of most people, like speech-to-text (Mukherjee et al., 2018) or language translation (Jiang & Lu, 2020) Some of the most common tasks that NLP is used for are sentiment analysis (Kastrati et al., 2021), text summarization (Rahul et al., 2020), and question-answering systems (Zhou et al., 2022), between many others. This shows how versatile the use of NLP can be as well as crucial to many industries, such as customer service, healthcare, and finance.

2.2. WORD EMBEDDINGS

For NLP, transforming words to a numerical form is a crucial step to use text data in models. One of the most popular ways to do this is with word embeddings, which is a representation of words in vector form in a continuous space (Incitti et al., 2023).

Before neural word embeddings there were the traditional methods for text representation that would rely on statistical and mathematical approaches to represent words in a numerical form. These methods are simpler when compared to neural network-based methods and produce sparser vectors. We will go through some of the simpler as due to being some of oldest and first methods, sometimes they are still used as benchmarks.

One of the most popular is the Bag of Words (BoW) technique as it is very simple (Harris, 1981). It treats a document as an unordered collection of words, disregarding grammar, and word order. The frequency of each word is what represents it, and the sequence of the words is ignored. Like this, the text becomes a sparse vector, where each dimension goes back to a word and the value signifies the occurrence or no occurrence of that word in the text. Despite being a quite simple technique, it is still used in multiple tasks. Its limitations are mostly the fact that it ignores semantic relationships.

Another immensely popular method is Term Frequency-Inverse Document Frequency (TF-IDF) that measures the importance of a word in a document relative to a collection of documents. So, it considers both the frequency and the rarity of a word on document in relation to the collection of documents. It is calculated by multiplying two parts. The first part is the Term

Frequency (TF) that measures how frequently a term appears in a document. While frequency is important, it cannot measure the importance of a word alone. That is where the second part of the multiplication enters. The Inverse Document Frequency (IDF) penalizes terms that are present too many times throughout the document set. It is calculated by taking the logarithm of the ratio between the total number of documents and the number of documents containing the term. Finally, by multiplying the TF and IDF values we get the final TF-IDF score. So, we have a weight assigned to each term in a document, that showcases its relevance to that document in the context of the entire document collection. Words with higher TF-IDF scores are considered more important, making them valuable. Some limitations of this method are the sensitivity to document length, as there is a tendency for the score to grow with document length, it also does not consider semantics.

In 2013, Mikolov et al. (2013) proposed a technique called word2vec that takes as input a text corpus and uses a neural network to create dense word embeddings representations of words as continuous vector space, where each word has a designated vector in a high-dimensional space. This technique can use one of two architectures: Continuous Bag of Words or Skip Gram, both of which use a sliding context window.

CBOW creates embeddings that use the influence of the context window on predicting the target word. Semantically similar words influence the probability in similar ways as they should be used in similar contexts. The Skip Gram method creates embeddings by using the target word to predict the context words surrounding it. Like this, semantically related words are expected to influence the prediction similarly, as they should share context.

Despite having a lot of success and popularity, Word2Vec still has limitations. It may struggle with words that have multiple meanings depending on the context. The model also may have a tough time when dealing with rare or out-of-vocabulary words, as it relies on pre-existing word embeddings, and it can struggle to generalize to unseen terms.

The GloVe (Global Vectors for Word Representation) was proposed by Pennington et al. (2014). This unsupervised learning algorithm maps the words in a space where the distance between the words is connected to how similar these words are semantically. It captures the probability of two words occurring together by factorizing the word co-occurrence matrix, this being what creates word vectors representations. GloVe can encode semantic relationships as well as syntactic relationships. As it depends on word co-occurrence, it may struggle with words that have their meaning being determined by the context.

One predominant model that leverages deep learning techniques to capture complex relationships between words and their contextual information is BERT. Although it is a transformer model, the embeddings of BERT can be used by other models as well. These embeddings have the context of each word in consideration in a bidirectional manner, meaning that they take before and after the word, providing great word embeddings.

The advantage of word embeddings over traditional methods is that they can capture semantical relationships between words, which helps capture subtle context and meaning. With Bag of words (BoW) or TF-IDF representations, two words that can be used interchangeably with the same intention have a similarity of zero as they are entirely different in writing, which is not ideal for most tasks. Also, when comparing BoW and TD-IDF with neural network-based word embeddings, the vectors created via neural network-based word embeddings can be considered low dimensional, due to being denser, which helps reduce computational complexity and makes it overall more efficient (Incitti et al., 2023). Neural network-based word embeddings also provide correspondence between semantic regularities and geometric properties, which means that relationships between words will have a geometric manifestation in the vector space. This is what allows models to extrapolate linguistic patterns. One of the most positive aspects is that pre-trained neural network-based word embeddings can be used for transfer learning. These will be models that are trained on large amounts of data and can later be fine-tuned for specific tasks.

A major drawback of using neural network-based word embeddings is their black-box nature, meaning that these vectors are noninterpretable as we cannot associate dimensions to the linguistic properties. They are also limited to their learned vocabulary and fixed representations (Arseniev-Koehler, 2022).

OpenAI released their own embedding model at the end of 2022, the text-embeddings-003, that substituted five of their other embedding models, as this model performed better than their older models and was able to do so across multiple domains. Longer context and smaller embedding size are also positive points for this model as well as a reduction in price (Greene et al., 2022).

2.3. CHATBOT

Chatbots were created with the objective of being able to simulate human conversation. The first well-known chatbot was ELIZA in 1966 by Joseph Weizenbaum at the Massachusetts Institute of Technology (MIT) (Weizenbaum, 1966). This chatbot had multiple objectives but the most known goal was to talk in a way that imitates a Rogerian psychiatrist, which is a non-directive, empathetic approach to counseling. ELIZA was based on simple pattern-matching techniques, it would analyze the inputs given to identify keywords and provide predefined responses based on the keywords identified. This made users engage in seemingly meaningful conversations. One of the more impressive aspects of ELIZA was that it was able to turn user statements into open-ended questions, which gave the illusion of comprehension and encouraged users to go deeper into their thoughts and feelings. ELIZA was not able to truly understand the user, but it captivated users and some even reported feeling an emotional connection with the chatbot, which surprised Weizenbaum.

Despite all of this, ELIZA was only able to mimic conversation based on its limited range of predefined answers. It lacked adaptability and versatility as it was incapable of evolving or learning new experiences (Shah et al., 2016).

In the 1990s, A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) was created by Wallace (2009). A.L.I.C.E. also utilized pattern-matching techniques but was made with the new AIML (Artificial Intelligence Markup Language), a scripting language that was specifically designed to create chatbot responses, to simulate dynamic and contextually aware conversations. The way A.L.I.C.E worked was by analyzing inputs and comparing them to a database of predefined patterns in AIML. After finding a match, A.L.I.C.E would use the corresponding response. This allowed it to cover a larger number of topics and increase the complexity of the conversations. A.L.I.C.E won the Loebner Prize in the years 2000, 2001, and 2004, due to its ability to be mistaken for a human. This competition awarded the most human-like computer programs.

However, A.L.I.C.E still had its conversational ability confined to predefined answers and lacked true understanding, which meant it give answers that were untrue and made no sense when working outside of its scope (Ramesh et al., 2017).

After A.L.I.C.E., chatbots saw significant improvements in the late 20th and early 21st centuries. NLP and machine learning technologies played a crucial role in enhancing the capabilities of chatbots. Smartphones and messaging apps made it much easier for chatbots to reach the average person. This made companies start using chatbots as their customer service agent, integrated in the customer service systems. Despite this increase in use, most of these chatbots still relied on rule-based systems and decision trees to help the customer with their queries, which made them extremely limited.

Deep learning methodologies came and enabled chatbots to go beyond simple predefined patterns and ways. Chatbots became able to adapt to the user queries by analyzing and understanding their inputs in real-time, while also being able to learn and improve the way they answered the user.

The introduction of virtual assistants, like Siri in 2011, made the step forward with chatbots that could answer questions and do tasks for the user, with the integration of the virtual assistant into other applications on the device. Along with the release of Siri, Google Assistant and Alexa also made chatbots more present in the daily life of the average person.

Nowadays, chatbots are mainly used for customer service, as companies use textual chatbots to answer customer queries and doubts and providing solutions to problems as quick and efficient possible (Daza et al., 2023).

There are multiple ways to classify chatbot, according to multiple characteristics. Following the classification of Adamopoulou & Moussiades (2020) chatbots can be classified according to:

- Knowledge Domain: there are open domain chatbots, that respond to general questions, and closed domain chatbots that only focus on a particular knowledge.
- Service Provided: We have interpersonal chatbots that just provide services to the user and live outside their personal lives and intrapersonal chatbot that also provide some company to the user and are used in the user's personal life. Inter-agent is a more recent type that provides communication between chatbots.
- Main objectives of the chatbot: There can be Informative chatbot, that as the name suggests, provide information as their primary goal, conversational chatbots that have as their primary goal to talk to the user like a human and task-based chatbots that provide services as their main objective.
- Input processing used for response generation method: Here there are three types of chatbots, rule-based, retrieval-based, and generative chatbots.
- Human-aid refers to how much help from the humans a chatbot gets, so if there is human intervention in one of the components of the chatbot, then it is human-aided.
- The built method refers to whether a chatbot is open source or not.

Looking forward, the future of chatbots is moving towards even greater and smoother integration into daily life, with ongoing research in areas like explainable AI, emotional intelligence, and multi-modal interactions. Chatbots are expected to become more seamless in understanding user intent, making them valuable companions in our increasingly digital and interconnected world.

2.4. Generative Pre-trained Transformer

The GPT (Generative Pre-trained Transformer) model is a state-of-the-art NLP architecture developed by OpenAI with the objective of diminishing the dependence of NLP in supervised learning, as it requires to have a large amount of labeled data to train a model.

The GPT models combine unsupervised training with supervised fine-tuning, which makes it highly adaptable to various tasks. The unsupervised phase of a GPT model involves training a multi-layer transformer. This has as its main goal to predict the most probable next token in a sequence, given the preceding context, with captures the complex patterns and long-range dependencies of language. Like this, the multi-layer transformer can handle both receiving and generating text inputs. After the unsupervised phase, the supervised phase uses label data of the target task or domain to fine-tune the model. This adapts the general language knowledge that the model got from the unsupervised task to a specific task or domain, making the GPT models very adaptable and versatile across NLP tasks. This adaptability prevails even in few and zero-shot scenarios, which means that the models can perform relatively well with only some examples of the task or even with no examples at all (Radford et al., 2018).

OpenAI introduced the first GPT model in 2018 (Radford et al., 2018). GPT-1, as it became known, has 117 million parameters, and showed exciting potential by generating coherent text. It was able to improve state-of-the-art results in 9 of 12 datasets in the GLUE benchmark, which is a collection of diverse natural language understanding tasks used to assess and compare language models' performance on a variety of NLP problems (Wang et al., 2018).

Released in 2019, GPT-2 is a significantly larger model with 1.5 billion parameters and had better language generation capabilities (Radford et al., 2019). GPT-2, which was built on the foundation of GPT-1, introduced significant design improvements. Layer normalization was placed at the input of each sub-block and added after the final self-attention block as well as the initialization process changed to accommodate the accumulation on the residual path with model depth. This was done by scaling the weights of the residual layers with a factor of a factor of $1/\sqrt{N}$, with N as the number of residual layers. The context size was increased from 512 to 1024, the batch size was increased to 512 and the vocabulary was expanded with 50 257 new words. For the unsupervised phase of the GPT-2, the authors built a large and diverse dataset by using a web crawler that would focus on the quality of the data to add the dataset used by the GPT-1. The GPT-2 model was able to zero-shot state-of-the-art performance in 7 of the 8 tested datasets by Radford et al. (2019), which shows its adaptability.

The GPT-3 was released in 2020 by Brown et al. (2020), and it was even bigger than GPT-2, with 175 billion parameters. The GPT-3 model is able to improve on the few-shot ability of the GPT-2, learning quickly with only some examples and then generalize this knowledge to unseen cases without more fine-tuning (Liu et al., 2021). Although, when fined-tuned to a specific task it can lose some of its generalization capacity, which is to be expected (Chowdhury & Haque, 2023).

The GPT-3 was optimized to create multiple models for specific tasks, like the code-davinci-002, that was fine-tuned on code generation. By optimizing the GPT-3 model to be a chat, OpenAI created the GPT-3.5-Turbo model, that is the base of ChatGPT (Kalyan, 2024). The GPT-4 model was released in March 2023, being able to receive text and images as inputs and has shown promising results.

Overall, all the GPT models share some of the same limitations. They are prone to have data bias, which means that they can reproduce stereotypes or misconceptions present in the data that is fed to them. On the same note, the GPT models are limited to the knowledge that they have, and although they have good generalization capabilities, it still only knows events and information that are present on its knowledge base. The GPT models can still display some inaccuracies due to ambiguity or lack of context understanding, although it improves much with the generations. They all also require a good amount of computational power as well due to their size and are expensive to train.

Multiple studies have been done on the different applications that the GPT model can have and their performance.

One example is the work done by Tsai et al., (2023) that created a GPT-based chatbot that answers questions about personalized pregnancy nutrition, NutritionBot. For the creation of the chatbot, a mixed approach was used. This involved a user-centered approach, where a persona was created, the collaboration with four medical specialists, to help with the design of the chatbot. A generic chatbot was created to compare with NutritionBot, which is integrated with ChatGPT and collects the patient's demographic information and generates recommendations based on that information. Since the data for responses and intents were implemented and defined by the research team for the generic chatbot, the chatbot was not flexible and may not comprehend questions which then leads to it not being able to provide the any feedback. This limitation could be solved by increasing the number of examples and the data that is fed to the chatbot, however this is a very costly process in terms of computational power. With the integration of ChatGPT, personalized nutrition recommendations can still be generated but with a much less expensive data training process. While the GPT-based chatbot showed responses that are well tailored to the input, it can still lead to biased, hallucinated, or misleading recommendations. Also, due to the model's nature, different responses will be generated each time, which could have varied different advice. This is problematic in the domain of health and medical recommendations.

2.5. Retrieval Augmented Generation

To fully understand what RAG is, we must start by understanding what retrieval is. Retrieval is the process of getting information or pre-existing responses from a predefined collection based on used input. This is called a non-parametric memory. So, if a chatbot is based on retrieval, that means that it is not generating responses but getting the content from a database (Zhu et al., 2021).

This type of memory use has positive aspects like efficiency, consistency, reduced training data, among others. However, it also has some cons as well, since it is very dependent on the training data, and chatbots that are retrieval-based tend to be inflexible and not understand ambiguity very well.

The alternative to non-parametric memory is parametric memory, where the knowledge that the chatbot has is stored in its parameters. This provides much better context understanding, handling ambiguity much better than non-parametric based chatbots. However, the expansion or correction of the knowledge base cannot be easily done as it requires retraining of the model, and these models are often computationally expensive to train, making it unviable to keep retraining.

To try and mitigate the negatives of each memory, there has been interest in creating hybrid techniques, which will combine non-parametric memory with parametric memory. Lewis et

al. (2020) proposed Retrieval-augmented generation (RAG) as a hybrid method that combines a pre-trained, parametric memory generation model with a non-parametric retriever. In their work, Lewis et al. (2020) uses a BART as their parametric memory and generation model and a dense vector index of Wikipedia as their non-parametric memory. The retriever used to access this non-parametric memory is a Dense Passage Retriever.

Lewis et al. (2020) created two versions of the RAG structure:

- **RAG Sequence**, where the same document retrieved will be used to create the full output sentence, bringing consistency to the process.
- **RAG Token**, where the document that is used to generate can differ at each token generation, bringing flexibility to the process.

RAG can be applied for different domains and industries. Among the most notable tasks that RAG has been applied to is open-domain Q&A (Mao et al., 2021), as well as specific domain Q&A (Siriwardhana et al., 2023), code generation and summarization (Liu et al., 2020; Parvez et al., 2021), and sentiment analysis (Zhang et al., 2023). RAG has also proven its benefits in chatbot tasks with numerous studies showcasing its utility in enhancing conversational AI.

Soong et al. (2023) implemented RAG for biomedical Q&A. For this study, a database of scientific papers related to diffuse large B-cell lymphoma disease (DLBCL) was created and these papers were pre-processed and split into segments of 4000 tokens. The embeddings used were provided by OpenAI's *text-embedding-ada-002*. The way the documents were chosen for retrieval was a similarity to the question ranking. The answer was generated in two phases. In the first phase, *text-davinci-003* was used to answer *k* times, each one using a different segment from the *k* retrieved, with some instructions tailored to minimize the inclusion of non-factual information into the text. Then, using these *k* answers, a final response was created by using *text-davinci-003* to summarize. This approach showed a that RAG LLM performed more accurately than an LLM alone.

Another study conducted by Ranjit et al. (2023) put RAG to the test by using it to generate Chest X-Ray reports, where once again, the reports generated by RAG were more based on the ground truth and factual in comparison to other methods. For this, Ranjit et al., 2023 built on a previously trained model (CXR-ReDonE), using this retrieval-based model as the non-parametric memory model. For the parametric memory, multiple models were assessed (*text-davinci-003*, *gpt-3.5-turbo* and *gpt-4*).

One of the most well-known problems with Generative AI is hallucinations, as many times the information provided is inaccurate, incomplete, or not in line with the prompt given. The original work on RAG (Lewis et al., 2020) compared BART with RAG with the BART model alone. By using RAG, the responses provided were more accurate and factual, which shows that RAG reduces hallucinations. This was further confirmed by Shuster et al. (2021) with their study

that concluded that the RAG method does reduce hallucinations by 60%. Although that is a particularly good improvement, it does not mean that the models do not hallucinate and still leaves many limitations unanswered such as not being able to recognize when it does not have the information required to answer a question (Chen et al., 2023).

3. RELATED WORK

The related work can be divided into two categories: Chatbots in Education, as this chatbot is being done for an academic environment and Chatbots in Organizations, as it will be a chatbot for Nova IMS.

3.1. CHATBOT IN ORGANIZATIONS

Organizations are always looking for ways to make their business processes more efficient and quicker. So, it is no surprise that there are so many organizations with interest in creating their own custom chatbot, due to its potential. A chatbot can help with parallelization of query answering and tasks, as one chatbot can interact with many users at the same time. A chatbot can work all day, every day, providing support whenever it is needed. When companies go ahead and make their own chatbots, the most common type of chatbot to be created is a customer service chatbot to help fulfill customers' inquiries. It makes sense to create these types of chatbots as many times customers' inquiries are repetitive and simple. But there is also potential for making chatbots where the end user is an employee of the company, as there are many tasks inside an organization that can be simplified with a chatbot.

One application of a chatbot can be to help with the recruitment of employees. When recruiting there are simple queries and questions that are repetitive to answer, like information about application status, rules of the company, among others. These simple and quick questions can be answered through a chatbot. There are also tasks that chatbots can do like find the best time to schedule reunions for the recruiters and summarize resumes to help pre-screen the candidates (Tadvi et al., 2020). Like this, chatbots help recruiters by diminishing their workload, allowing them to focus on the more complex questions that applicants might have as well as focus on preparing the recruitment process.

A chatbot can also help train new employees, as it can help with the learning process that might come from getting a new job in a new company (Casillo et al., 2020). Often, there are required courses, documents, and information a new employee reviews when they enter a company, like policies or how the working day processes work. A chatbot can break down content for the new employee and answer questions regarding it, rather than having the new employee solely dependent on their mentor or the human resources department.

Another application of chatbots by organizations is creating a chatbot that solves tasks for the employees and helps with connecting them inside the organization (Frommert et al., 2018), like a chatbot that is able to schedule reunions and book rooms that are available for them. This saves much time as the employees do not have to compare calendars to find an opening that works for both of them and also do not have to contact the administrative services to figure out which rooms are available for booking.

KlimaKarl is an example of a chatbot that is used to improve communication in the office, as it was created to inform the employees of a company on how to behave in a more conscious way in relation to the climate in office and in everyday life (Hillebrand & Johannsen, 2021).

There is also the option of creating a question-and-answer chatbot for employees, like the ones used for customer service question-and-answer. It is common that employees have questions about the organization they work for, like where is a department located or what are their benefits or by creating a task-based chatbot. A study done by Fiore et al. (2019) shows that there is interest by the employees in having an internal chatbot that would support some of their queries and questions. And while there are multiple studies documenting how the GPT models can be applied to make helpful chatbots for enterprises (Ayinde et al., 2023), the challenges that there are with building one, or how open the companies are at the idea of building one (George et al., 2023), there are not many published papers on chatbots created specifically for supporting one's organization community, having them as the primary users of this technology.

Many organizations, however, report using the GPT technology to make their own chatbots for customer service or even to offer a service. The AMA (Agency Administrative Modernization) presented a chatbot that can receive and output voice and text to answer questions about the Digital Mobile Key (Microsoft Portugal News Center, 2023). CTT also released Helena, a customer service chatbot based on the GPT technology (CTT, 2023). Both projects use Azure OpenAI as its main component. Other companies, like Snapchat with MyAI (Snapchat, 2024), Duolingo with Duolingo Max (Duolingo Team, 2023) and Quizlet with Q-Chat (Quizlet, 2023), used the GPT model to create tools and products to improve consumer experience.

While not focusing on any company or a chatbot for a specific need, a work done by Jeong (2023) demonstrated implementation methods and tools for applying LLM models with RAG for business domains and enterprises, while exploring how to optimize these techniques and confirming the benefits and how practical is the RAG model.

The implementation of a chatbot in an organizational setting comes with some challenges that need to be addressed in order to have the chatbot work efficiently and overall be successful (George et al., 2023). The first obstacle that many organizations will face is integrating said chatbot into their systems and making the chatbot work with the workflow. Each organization uses a set of software applications, often unique to the organization as they were adapted for the enterprise's needs. So, to make a chatbot work in a seamless way with this software can require a great amount of planning and customization, which can take an exceedingly long time.

With this, developers also must make sure that the chatbot has the information to answer complex and industry specific questions that demand a particularly good understanding of the subject matter. This must be paired with maintenance and regular updates of the chatbot in

order to stay effective and up to date. This can be a tricky thing to do, as there are constant evolutions and new information coming out in the business world. Customers and staff might expect the chatbot to answer according to these new developments. So regular updates are necessary to ensure that chatbots remain aligned with the latest industry trends, regulatory changes, and company-specific updates. If not kept up to date, the chatbot's results may have inaccuracies, which can lead to frustration among users. While fine-tuning and RAG help with both the creation of a domain specific chatbot and the update and maintenance of the that domain, they are not methods without faults, and it is important to keep that in mind. The scope of the chatbot has to be well defined for it to not leak or do anything the enterprise might not want it to do. Recently, both DPD (Quiroz-Gutierrez, 2024) and Chevrolet (Bastian, 2023) released chatbots that were not properly defined and safe-railed, which led to the chatbots generating responses that were not adequate.

But none of this matter if the target users do not use the chatbot regularly or at all (Brachten et al., 2021). People can be resistant to change and might have a negative bias about using a technology that relies on artificial intelligence. So, it is important to be clear about what this technology is and how it can help. Organizations might release some training and some support to the employees so that they can feel comfortable and confident in the use of chatbots for their daily tasks and processes. There might also be concerns about job posts being replaced by technology, and these should be promptly addressed by the organization to let the employee anxiety reduce.

Data security and privacy continues to be a challenge with chatbots in organizational settings, as it is important to that the users, whether these are customer or the workforce, can trust the chatbot with their information as well as it important that the chatbot does not leak any sensitive enterprise information in the cases where the chatbot is directly connected to the organization's database.

Lastly, there are always ethical concerns with the deployment of a chatbot, which cannot be overlooked. It is the organization's responsibility to make sure that the chatbot meets the ethical guidelines and respects the user.

3.2. CHATBOT IN EDUCATION

When talking about chatbots in education, most of the conversation focuses on chatbots for learning or in chatbots for answering frequent questions.

Universities benefit from using question answering chatbots as they can be used to reduce the workload of the administrative offices' workers. Administrative offices can receive a substantial number of questions and inquiries per day, and even sometimes the same question is asked multiple times by different people (Lee et al., 2019). Most of these questions and inquiries are very simple and can be answered by a chatbot, which means that the person with the question can get faster, whenever they want responses, and the staff of the

administrative offices have more time to dedicate to the questions and inquiries that cannot be solved by the chatbot and really need their attention.

One example of this type is in the work done by Ranoliya et al. (2017), that created a chatbot that would answer the most asked questions done by students of Manipal University. The chatbot was done with AIML language, which means that the chatbot created is limited to a set of manually created rules, hence having difficulties with adaptability, and understanding more complex approaches to chatbots. A more recent and advanced example would be Aisha, a chatbot leverages the capabilities of the GPT-3.5 to create a chatbot for Zayed University Library (Lappalainen & Narayanan, 2023). There is also EduChat that a hybrid chatbot that uses rule-based methods, with a random-forest intent classifier and the GPT-3 model to answer student's questions (Barkalov et al., 2023). Chatbots tasked with helping students go through universities admissions and applications are also quite common. Day & Shaw (2021) created a GPT-2 based chatbot that pretrained with the encyclopedia question-and-answer (BKQA) dataset and fine-tuned with the Tamkang University Admissions question-and-answer (TKUAQA) dataset to help with university admissions. There is also DINA that provides the same service but for the Universitas Dian Nuswantoro (Agus Santoso et al., 2018).

Although these chatbots support staff by reducing their workload and streamlining their operations, none of these chatbots has as the target user the university staff. There are no papers that report on the implementation of a chatbot like that, however, the University of Canberra reports having the chatbot pair Lucy and Bruce, where Lucy helps the students with their queries and Bruce helps the staff (Perry, 2018).

Another way that chatbots can be applied in education is using them as an auxiliary tool for teaching and learning. The utilization of chatbots for this purpose started in the early 1970s (Amin Kuhail et al., 2023), so multiple studies have been conducted throughout the years evaluating how well can a chatbot be used as a tool for teaching, as well as a tool for learning.

The use of chatbots for teaching and learning is extremely attractive since it can help with customization of the learning process and can aid teachers in answering the students' questions. The number of students a class has can be very large, as many universities are giving lectures at a large scale and the rise of massive open online courses (MOOCs), and with that can be hard and overwhelming for the teachers to provide individual and personalized support for each student. A chatbot that has knowledge on how the class works and/or the subject of the class can be a first point of contact to answer questions and help with tasks, and if after using the chatbot a student still needs help, then the student can contact the teacher. This type of chatbots can be used to provide feedback on how the students are doing in a class, to improve the performance of students that might be struggling with a subject or even as motivational agents (Winkler & Soellner, 2018).

A common subject where teaching/learning chatbots are applied is language learning. As customization is particularly important to language learning since each person will have

different levels of fluency and different areas where they are stronger at. For example, one person might be very good at grammar and have a hard time writing due to poor vocabulary, while the next one can be the reverse. That is a key benefit that chatbots bring, as the student can choose what they want to practice. Another way that chatbots help is with accessibility, as language learning needs continuous practice for the student to be able to make progress and with a chatbot students can study and practice on the go, wherever and whenever they want. Moreover, these chatbots offer a dynamic and immersive learning environment. Learners may improve their speaking, listening, and understanding abilities by engaging in interactive discussions that replicate real-life conversations. Furthermore, language learning chatbots can integrate multimedia elements, such as images, audio, and video, to enhance the overall learning experience (Huang et al., 2022).

The implementation of chatbots in an educational setting has some challenges (Okonkwo & Ade-Ibijola, 2021). Although the advancements in teaching are positive, when compared to a human teacher, GPT-3 model is still lacking on being able to speak, understand the student and help the student like a teacher (Tack & Piech, 2022). There are also ethical considerations with how much people trust the chatbot, how transparent the chatbot is with how it works, and privacy (Adiguzel et al., 2023). Most of these challenges that chatbots in organizations face are also appliable to chatbots in education and academic settings.

4. METHODOLOGY

This section presents the methodology adopted during this project. While there it was not applied a formal methodology, inspiration was taken from the CRISP-DM, a widely used and popular methodology for data science processes (Wirth & Hipp, 2000). This methodology is composed of 6 different phases, that tend to be arranged in a cyclic manner, with the output of the previous phase being fed to the following phase. Despite the tendency to be arranged in a cyclic manner, the order of these phases is not strict and there is room for moving back and forth between the different phases.

Inspired by the CRISP-DM methodology, the methodology of this project has 6 phases. The first 3 phases were taken directly from the CRISP-DM methodology, while the fourth phase was adapted to a Design and Development phase and the fifth phase was also adapted to a Testing and Evaluation phase. The last phase, Deployment, was taken from the CRISP-DM methodology as the first three were.

The first phase of this process is the Business Understanding phase. This phase focuses on understanding the goals and objectives of the project as well as the requirements from a business perspective. With these set up, a preliminary plan can be developed to solve the problem defined.

The following section is the Data Understanding phase. In this phase the goal is to collect the data and as well as describe it. This phase is also when the exploratory analysis is done, enabling familiarity with the data, identifying possible problems and steps that will need to be taken in the next phase. Following the Data Understanding phase is the Data preparation phase, that covers the steps needed to transform the data from the raw initial stage to being ready for the modeling phase.

The fourth phase is the Design and Development of the chatbot. This involves the design of the architecture used for the chatbot and the definition of the functionality of the chatbot as well as the implementation of said architecture.

The fifth phase of this methodology, the Testing and Evaluation phase is where the functionality and performance of the chatbot will be evaluated, to check if the chatbot behaves as expected and meets the business goals.

Finally, the last phase of this methodology is the Deployment, that involves the planning of the deployment of the model, as well as the monitoring and maintenance of the model. This phase is out of the scope of this work as addressed in the Limitations and Future works section.

4.1. BUSINESS UNDERSTANDING

As previously stated, this phase wants to determine business objectives, assess what resources are available and what are the principal requirements, to define the goals of this project from a technical perspective. The goals of this project were already presented in the introduction as well as the theorical background being presented in the previous chapter.

There are numerous administrative processes, student services, and vast course options in academic and organizational settings. Students, professors, and administrative staff may find it difficult to navigate through this enormous amount of information and quickly discover the one they are seeking for. And given the rapid growth of technology, there is a growing need to make use of the newest technological innovations to boost access to institutional information in Nova IMS. In this context, chatbots are a great answer, with the most recent technological innovations showing that the ones powered by LLMs as the most advanced and appropriate, especially the GPT model. Chatbots offer a unique approach to delivering information, providing immediate responses to inquiries, and helping users through administrative and academic procedures. This is especially important in a university context where fast and accurate information is essential for decision-making and enhancing the overall educational experience.

With a chatbot Nova IMS can reduce workload of human resources, as it can automate responses to frequently asked questions (FAQs) and routine inquiries from students, faculty, and administrative staff. A chatbot will also improve accessibility as it provides 24/7 access to information regarding courses, programs, administrative procedures, and campus facilities, ensuring that users can obtain information promptly and conveniently. It will also enhance the user experience, as offering a user-friendly and dynamic interface allows seamless interaction for various queries, thereby improving overall satisfaction among users

However, the use of the generic GPT models in specific context such as Nova IMS' has inherent limitations. While these models can generate contextually coherent text, their effectiveness, and the knowledge they provide is linked to their training data. Generic models often lack the specialized knowledge necessary to effectively navigate the specific intricacies and requirements student and staff needs at Nova IMS.

Therefore, there is a need for a tailored solution combining the general capabilities of LLMs with a deep understanding of the unique domain of Nova IMS by using the RAG framework.

In conclusion, there is a need for an advanced, customized chatbot to address the specific challenges posed by the very specific domain of Nova IMS. Using an GPT model along with RAG ensures accurate and contextually relevant responses as well as supporting the diverse needs of students, faculty, and administrative staff.

4.2. Data Understanding and Preparation

Right now, if we want to know something specific about Nova IMS, there is the option of either emailing/calling the different services available or going to the Nova IMS website to try and figure out the answer to our question ourselves. The Nova IMS website is a very good source of information about the institution, so it was chosen as the data source for this project. However, having to manually gather this data from the website is a very tedious and arduous task. To make the data gathering process more efficient and quicker, the data for this study was gathered by web scraping the Nova IMS, more specifically the "Ensino" branch of the website.

Figure 4 showcases a general list of pages of the **Education** branch. As it can be seen, it contains all the information about what the academic options of Nova IMS are, between bachelors, post graduate and master's programs, a doctoral program as well as smaller courses such as executive education, workshops and short-duration courses and micro-accreditations in Hospitality & Tourism. It has specific information on what each of these academic options are focused on, what is their study plan, and some testimonials of previous or current students that took that academic path. In the pages about each option there is also information about the tuition, scholarships, and some frequently asked questions. This branch also contains information about Erasmus and study abroad options and how can a student study abroad as well as information for the students that are interested in studying in Nova IMS during exchange program experience or being a student from Nova IMS while having not study in Portugal. It also has information about mentoring and employability, as well as companies that partner with Nova IMS for internships. It is also in this branch that a student can find information about scholarships, the social services department of the university, insurance, the student association and the prizes and the honor roll.

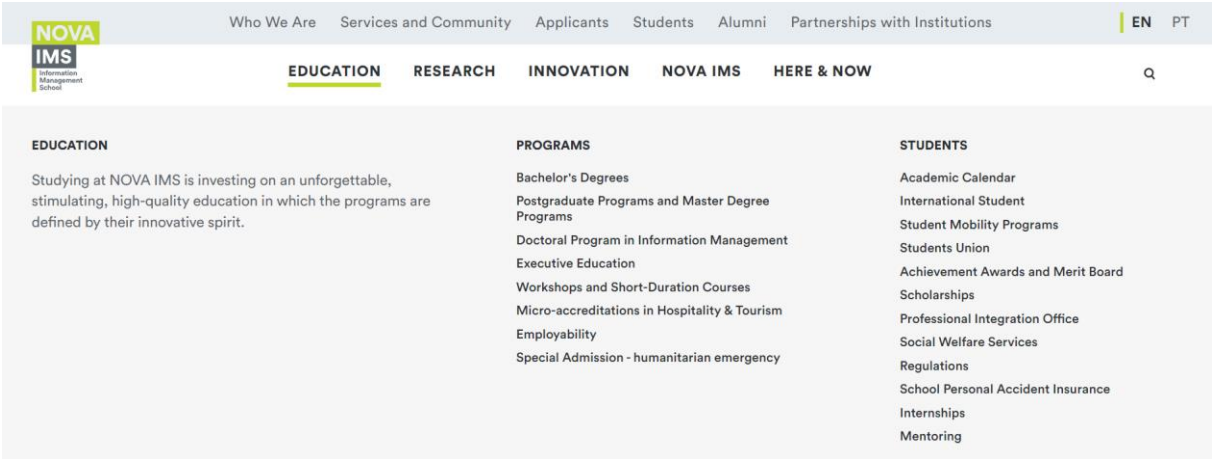

Figure 4 - Menu of the Education Branch of the nova IMS website

During the data gathering, there were 287 txt files extracted from the website, most in Portuguese, as the ones directed at international students are in English. In order to obtain these files, web crawling and web scraping was applied, starting from the page titled “Ensino”. As the pages of this branch are not very consistent in shape, it was very hard to fully customize the way the scraping worked. A more general way was used to extract the text, that was cleaned up after. The web scraper got only the text located in the main element of the page, as the footer and the header are repeated in each page and only the content specific to that page is relevant information. The text present on each page is exported in a text file, so each page gets their own file. The web crawling was applied to try and get all the pages in the “Ensino” branch of the Nova IMS site, so in each page, the links that started with *https://www/novaims/unl/pt/pt/ensino/* would be extracted to a list to be scraped next.

Before the cleaning of the data, the word count of each file was done, and the general statistics of this variable were explored, as seen in Table 1. There are some documents that are small, and most documents have less than 1564. However, there seem to be some large documents too, with one as big as 8278 words.

Table 1 - Word Count Statistics before cleaning the Data

	Word Count before Cleaning
Count	287
Mean	1411,10
Standard Deviation	1765,80
Min	5
25%	259
50%	504
75%	1564
Max	8278

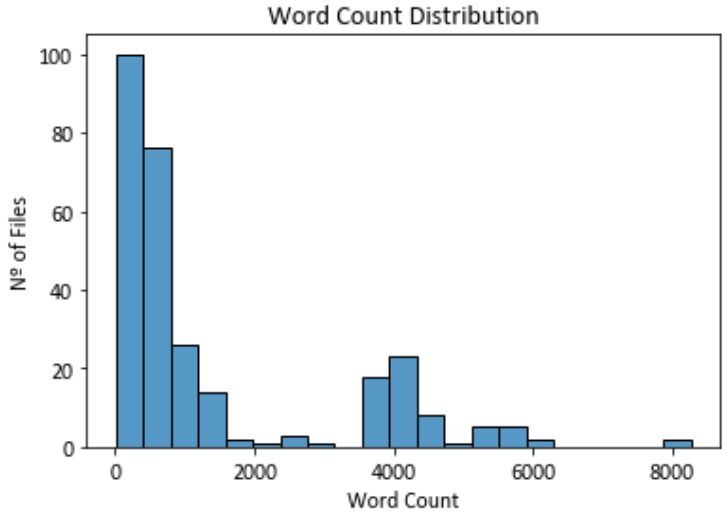

Figure 5 - Word Count Distribution Before Cleaning

To better visualize the distribution of the word count variable, a histogram was made, as it can be seen in Figure 5. The word count before histogram supports the skewedness towards smaller text and shows that the big document is an outlier between the others.

Table 2 has can be seen the 20 most common words, with word cloud in Figure 6 to complement it visually. These consist mostly of stopwords, which are words that occur frequently but carry very little meaningful information or semantic content. Between the other words, words like 'professor', 'optativa', 'curso' and 'convidado' are expected due to the academic and curricular themes, as the documents have information about subjects, classes, and curricular plan. After a quick look at the documents, 'ilhas' and 'república' and other geographical terms present in the word cloud are due to input boxes that are in multiple pages of the documents. The word “nova” is likely due to the name of the faculty. Finally, the words 'saber' and 'mais' together showcase that there are a lot of hyperlinks for other pages on the website pages that are recorded as that text in the documents. In total, these 20 most common words represent 25,38% of the words in the documents, which is a very significant part of the data.

Table 2 - Most common words in the dataset

	Frequency	Percentage (%)
de	23386	5,84
e	10571	2,64
a	9251	2,31
da	6753	1,69
o	6617	1,65
do	5597	1,40
em	4429	1,11
mais	4255	1,07
que	3318	0,83
para	3315	0,83
saber	3224	0,81
ilhas	2958	0,74
no	2484	0,62
os	2420	0,60
república	2309	0,58
nova	2304	0,58
curso	2196	0,55
convidado	2148	0,54
professor	1987	0,50
optativa	1969	0,49

Figure 6 - Word Cloud for the documents before cleaning

However, when working with LLMs, the preprocessing of the text is more flexible and does not require these techniques. While the goal of removing non-relevant, or less relevant text is still the same, the way it was applied was by going through the documents to figure out which parts of the text might be confusing or repetitive for the chatbot.

First, text that was duplicated across documents or documents that were duplicated were removed, as well as pages that were empty, which lowered the document count to 194. Some documents were also separated in two, as it made more sense for them to be in two. Some information from other pages of the Nova IMS website was added to the documents to make them completer and more useful. Information that was in image format was also added to the documents.

Words that would indicate the presence of a button or a link to another page were removed, like “Saber mais” or “Clique aqui”, as this is not relevant, and it implies that there is information missing. Menus and text that belonged to input boxes also had to be removed. The title of each page had to be added manually to many documents as well. It was also checked if the format of the text was correct, and the information and numbers were aligned correctively with the text.

After the cleaning of the data, the word count of each file was again explored, as seen in Table 3. While there are still small documents, they are not as small as before, and now most documents have less than 2151. The larger documents seem to be less and the largest now has 4468 words, almost half of the one that was previously the largest.

Table 3 - Word Count statistics after cleaning the data

	Word Count after Cleaning
Count	194
Mean	976,32
Standard Deviation	1096,20
Min	21
25%	226
50%	372
75%	2150,75
Max	4468

Once again, to better visualize the distribution of the word count variable, Figure 7 has a histogram was made for the word count variable after cleaning. The word count after still shows a skewness smaller text although less prevalent and now there are no outlier documents that are very large.

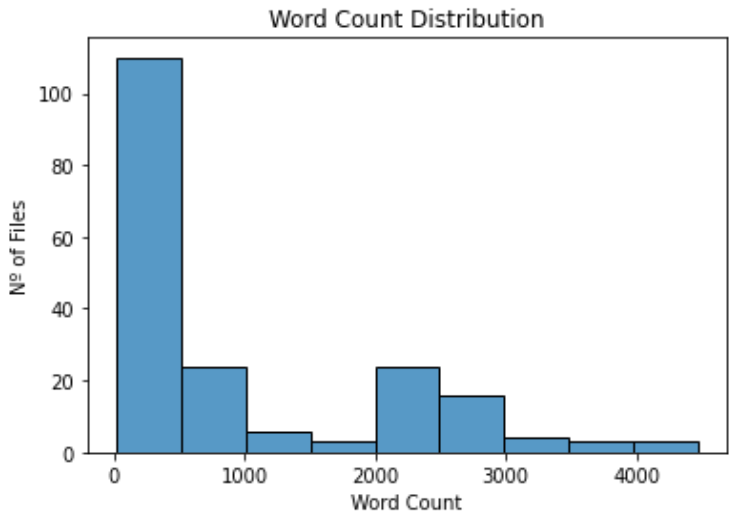

Figure 7 - Word Count Distribution after cleaning

The analysis of the 20 most common words was also done for the documents after cleaning, and it still mainly consisted of stopwords, similarly to what was shown before. Interestingly, the word nova is not in the top 20 anymore, it is the 21st most common word, with 1094 occurrences. The top 20 words now represent an even bigger part of the dataset, together with 29.40% of the words present in the documents.

4.3. Design and Development

To create the chatbot, Microsoft Azure OpenAI service is used. This service is a collaboration between Microsoft Azure and OpenAI, giving developers the ability to use the OpenAI models, like GPT-3.5-Turbo or Embedding models. The pros of using this service over OpenAI's service are that with Microsoft Azure there are security capabilities, like private networking, regional availability, and responsible AI content filtering (mrbullwinkle et al., 2024).

The main difference between gpt-35-turbo and gpt-35-turbo-16k is that the context window of the gpt-35-turbo is much smaller than gpt-35-turbo-16k, making the gpt-35-turbo cheaper too. For this work, a bigger context window is important so the gpt-35-turbo-16k was used.

To use this service is important to know what tokens are. Azure OpenAI divides text into tokens, that can be words or chunks of characters. Table 4 shows the models available for deployment and their token limitations. It is also important to define that a prompt is the text that the users send to the model, while the completion is the model answer to the prompt.

Table 4 - Models Available

Model Name	Tokens per min	Requests per minute
text-embedding-ada-002	6 000	36
gpt-35-turbo	10 000	60
gpt-35-turbo-16k	4 000	24
text-embedding-3-small	4 000	24

By using the work done by Jeong (2023) as a guideline for building this chatbot, the LangChain framework was adopted. LangChain is an open-source framework made specifically for the development of applications based on languages models, allowing for the introduction of context awareness of a specific domain to the model (Langchain, 2024). Like this, LangChain is used as a bridge that allows for interactions with multiple LLMs, with a large number of tools that allow for prompt creation and handling. LangChain is also helpful for conversational memory creation, enabling the management of past chats and queries, crucial to chatbots. There are also index tools for better organization of documents, as well as an easier way to chain multiple LLMs in order to respond to more complex tasks. Lastly, LangChain provides tools to apply RAG on language models, including document loaders, text splitting, text embeddings, vector stores, retrievers, and indexing.

To be able to use the documents, they must be stored somewhere. However, traditional relational database management systems are not recommended to store such unstructured data. To answer this need, there are Vector databases, that are made specifically to store the

embeddings of the documents in an efficient manner and provide fast and easy access to the documentation. Like this, retrieval can be performed with similarity search.

With data presented as vector embeddings, it is possible to identify similarities among different parts of the data and retrieve data that is comparable to a certain embedding. The query is initially transformed to embeddings by using an embedding model, and after which the Vector Store receives this vector embedding and runs a similarity search against the other embeddings stored on the database, retrieving all pertinent data. These valuable vector embeddings are then sent to the LLM, that uses this information to construct the response to the user query.

For the vector database of this work, Chroma was used due to its easy implementation and compatibility with LangChain. Chroma is also a free and open-source project, making it open to suggestions, improvement and quick issue solving (Chroma, 2024). At the moment, Chroma only allows for local storage, although the creation of a hosting service is in the future plans. Chroma has 4 functions in the API, so it is simple and easy to understand.

To create the chatbot there are several steps that need to be taken. Figure 8 shows how the different elements of the chatbot come together to answer a prompt.

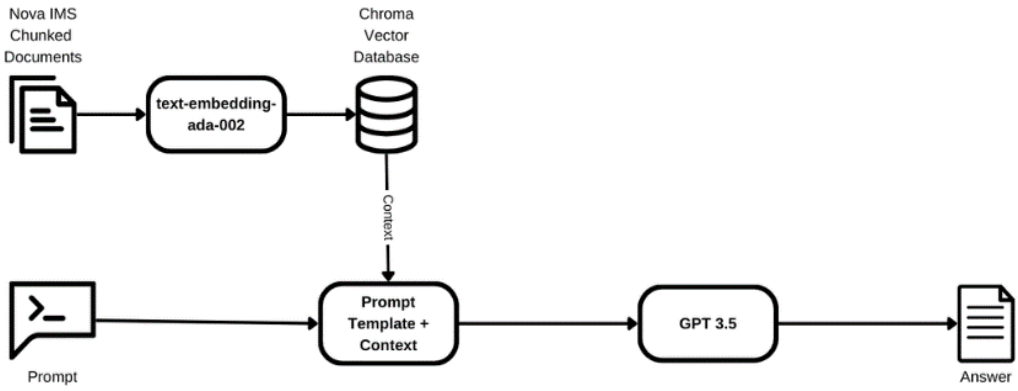

Figure 8 - Chatbot design

The chatbot has the domain information regarding Nova IMS stored in a vector database in embedding form. The chatbot can also be instructed by using natural language. When the user asks a question to the chatbot, it transforms this question into an embedding and using a similarity measure, it pulls the top k most related chunked documents. These documents are fed to the GPT model along with the prompt given by the user and an instructional prompt. The GPT model uses the information to generate an adequate response to the query. All of this is stored in a memory, so that during a conversation, if the user needs to do another query that is still related to the previous query, the GPT model is able to respond, as it is fed this memory, allowing for it to have context about the previous query.

Jeong (2023) has based the implementation of RAG on the following steps:

1. Source Data Collection and Extraction: this is where structure and unstructured data is gathered and collected as well as material that is defined the rules of the chatbot. This step refers to section 4.2.
2. Chunk Generation: during this step, the text is split into smaller units, known as chunks. This helps retrieval work better and helps the model as it will receive only the most relevant parts of a document.
3. Embeddings: the chunks get transformed into numerical vectors.
4. Building the Vector Database: this database will store the embeddings, facilitating the search and the calculation of similarity measures.
5. Integration of Prompt and Search Results: this is the retrieval part, where, based on the prompt, relevant information is searched for. The retrieved chunks are sent to the LLM for the answer generation.
6. Answer Generation: with the retrieved information as a basis, a response text to the user query is generated. The type, length, and linguistic style of the generated text can be specified with prompt instructions.

Following the previous steps, the chunk generation should be performed. The data gathered was loaded with a directory loader and then chunked into both 500-character and 400-character sized pieces so that retrieval is able to be performed in easy and quick manner. It was also stored with different overlapping percentages for each size: no overlapping, 10% and 20%. In total, 6 different vector databases were created with different configurations so that the best one can be evaluated for.

This preprocessed, chucked data was transformed into a numerical vector by the Azure OpenAI model, *text-embedding-ada-002*, which corresponds to step 3. Into step 4, the Vector Store is created with ChromaDB in a persistent local memory. As there is a limit on how many tokens the embedding model can be fed per minute, the embeddings and storing of them had to be done in batches.

Based on step 5, two retrievers, one based on cosine similarity, and another based on MMR similarity, were initialized to find the top 4 relevant chunks related to the query. The number 4 was chosen as it is the maximum of documents that were need when creating the test dataset and more might be too many documents, given the context window.

As it was essential that the chatbot was able to respond in a conversational manner, a *ConversationalRetrievalChain* was created so that a chat history was created and fed to the model, making the chatbot able to remember the previous query in a conversation and respond accordingly if need.

All these steps are organized and activated together in a loop, that receives a user prompt and answers it. This loop only stops when the user asks to with “encerrar” or “quit”.

4.4. Evaluation Metrics

The evaluation of chatbots can be very hard, as there are many aspects to be checked. Chatbots can be evaluated on whether they provide accurate information, on how well written that information is, how relevant that information is in relation to the query or the conversational chain and how good they are at during the experience. To add to the multitude of way that chatbots can be evaluated there is also some subjectivity to most of these evaluations, for example, if I ask about the education offering of Nova IMS, one can think that the only relevant information to this question is the name of the academic courses, while another might think this is incomplete and it should also include a small description. There is also the problem of variety of the chatbot answers, as the chatbot will not always answer in the same way, so sometimes it might provide better answers than other times. To try and deal with this, the chatbot was evaluated in different ways.

To add to the difficulty of evaluation, the evaluation for this work is on a very specific topic, so there is no ready to use dataset of questions-answers that can be used to check if the chatbot is accurate. Because of this, datasets had to be created with common questions that universities offer. With these as basis, the questions were tailored to the Nova IMS case and responses were created.

To assess if the retrieval is accurate and which measure should be used on the retriever to choose the documents, the documents with the information answering the dataset questions are also added to the dataset as another variable.

The resulting two question-answer datasets were the following:

- The first dataset is composed of 30 questions that were chosen to cover most of the topics that were in the context provided to the chatbot, referred as **30qna_dataset** for the rest of this work. The questions of this dataset are on Appendix A. These questions also varied in complexity and tried to mimic a real interaction with the chatbot. This first dataset will be used to evaluate the retrieval measure, the chunking and overlapping of the documents and the overall capabilities of the chatbot to answer questions within the new context provided.
- The second question-answer dataset was created to evaluate how well the chatbot responds to the prompt template and to tailor it better as well, so it is composed of questions that are not related to Nova IMS or that are but are not on the context provided. This dataset will be referred to as **10qna_dataset** for the rest of this work. It is expected that the chatbot does not respond to the first type and say that it only answers to questions related to Nova IMS and to the second type, the chatbot should provide an alternative way to get information since it does not have that information.

Various aspects of the chatbot components were evaluated, in order to determine the optimal configuration for its components. As stated in the previous paragraphs, the retriever similarity measure needed to be assessed to choose the most accurate one. For the vector databases,

there is the need to chunk the documents, as the context window is limited to 16k of tokens, so the chucking size and the overlapping parameter were evaluated as well. As a prompt template will be added to the chatbot, for better customization, there was also the need to evaluate different prompt templates to get to the best one. A last manual verification of the answers was done to assess the general quality of the answers and the hallucinations that the chatbot would provide. For the evaluation, the RAGAS framework was used.

The RAGAS framework was created to evaluate RAG implementations on LLMs (Es et al., 2023). This framework has multiple applications: it allows for synthetic generation of a test dataset to be used in evaluation; use LLM-assisted evaluation metrics on the performance of the RAG application; it helps monitoring the quality of the applications in production with smaller and cheaper models, while still providing valuable insights into the performance of the applications. Like this, this framework is an important foundation to those wanting to implement RAG with their LLM applications.

The metrics that RAGAS provides can be divided into component-wise evaluation metrics and end-to-end evaluation metrics. Component-wise metrics evaluate both the LLM and the RAG component individually, while end-to-end focus in evaluating the overall performance of the pipeline, allowing to assess the user experience.

For the component-wise metrics, there are 6 in total, 2 for evaluating the LLM and 4 for evaluating the RAG component of the application. The 2 that evaluate the LLM component are:

- Faithfulness, which has as goal to evaluate how factually accurate the answer generated by the LLM model. It is measured between 0 and 1 and the higher the score the better. This variable is calculated by first obtaining the number of claims that are in the generated answer and then number of those that can be inferred from the given context provided by the retrieval. Then the formula is as follows:

$$\text{Faithfulness} = \frac{|\text{Number of claims that can be inferred from context}|}{|\text{Total number of claims in the generated answer}|}$$

- Answer Relevance, that evaluates how important the generated answer is to the question that was given as prompt. It is also a variable where the higher the score the better, meaning that the higher the score the more relevant the answer is to the query, which in this case means that the answer is complete and does not contain redundant details. To calculate this measure, an LLM is prompted to generate questions from the generated answer and the mean cosine similarity of between these questions and the original question is calculated. Having E_{g_i} as the embedding of a generated question i, E_o is the embedding of the original question and N is the number of generated questions (default is 3), the formula goes as follows:

$$\text{Answer Relevancy} = \frac{1}{\text{N}} \sum_{\text{i=1}}^{\text{N}} cos(E_{g_i}, E_o)$$

For the retriever component, there are the following 4 metrics:

- Context Precision evaluates if the contexts that were retrieved by the retriever are needed to get to the ground truth answer. This is done by calculating the relevancy for each chunk on the retrieved context to the ground truth answer.
- Context Relevancy evaluates whether the context chunks that are retrieved are relevant to the question. It ranges from 0 to 1, having better scores meaning that the chunks are relevant to the question. To get this metric, the number of sentences in the context that are relevant to the question ($|S|$) is calculated divided by the total number of sentences in the context. The formula goes as follows:

$$\text{Context Relevancy} = \frac{|S|}{|\text{Total number of sentences in retrieved context}|}$$

- Context Recall is a metric that evaluates how well the retrieved context aligns with the ground truth. The values range from 0 to 1, with higher values meaning better performance. To calculate this measure, each sentence of the ground truth is analyzed to see if what it says can be attributed to the context retrieved. Like this the formula goes as follows:

$$\text{Context Recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

- Context Entities Recall measures the recall of the retrieved context in relation to the entities. It first finds the entities present in the ground truth (GE) and then the ones present in the retrieved context (CE). Then, it takes the number of elements in that are in the intersection of these two groups and divide it by the number present in the GE , as given by the following formula:

$$\text{Context Entities Recall} = \frac{|CE \cap GE|}{|GE|}$$

For the end-to-end evaluation metrics, there are 2 in total. These are:

- Answer Semantic Similarity is one of the most commonly used metrics in evaluation of LLMs. This measure assesses if the generated answer is semantically similar to the ground truth, having the values fall between 0 and 1. For this measure to be calculated, both the ground truth and the generated answer are put through an embedding model to get their embeddings. After this the cosine similarity is calculated between both vectors.

- Answer Correctness aims to calculate the accuracy of the generated answer in comparation with the ground truth. It does this by calculation the factual overlap between the ground truth and generated answer, with true positives being statements that are present in both, false positives statements that are only present in the generated answer and false negatives statements that are only present in the ground truth answer. With these calculated, it calculates the F1 Score as follows:

$$F1\ score = \frac{|TP|}{\left(|TP| + 0{,}5 \times (|FP| + |FN|)\right)}$$

  And then it uses the Answer Semantic Similarity measure and calculates a weighted average of the Semantic Similarity measure and the F1 Score. The default weights being [0.75, 0.25].

For this project, 6 of the measures were used, the 2 metrics that evaluate the LLM component, Faithfulness and Answer Relevance, 2 that evaluate the RAG component, Context Relevancy and Context Recall and the 2 for the end-to-end solution, Answer Semantic Similarity and Answer Correctness. The way these will be applied will be further explained in the following section.

5. RESULTS AND DISCUSSION

The results and discussion will present the numbers obtained after that implementation of the methodology previously presented, while also discussing its impact. Both the retrieval method and the chatbot were evaluated with different parameters to achieve the best possible result. This section starts by presenting the steps and datasets used in the evaluation, in section 5.1. Then each of the following sections (5.2, 5.3, 5.4, 5.5) corresponds to one step of the evaluation procedure.

5.1. EXPERIMENTAL SETUP

As previously stated, evaluating a chatbot can be hard as there are multiple aspects of it that need to be evaluated. The testing of this chatbot tried to be as comprehensive as possible and evaluate multiple aspects. As the chatbot is composed of multiple parts, it was also necessary to evaluate each of them.

As previously stated for the evaluation of this chatbot, two question-answer datasets were created, **30qna_dataset** and **10qna_dataset**, that will be used in different parts of the evaluation.

As multiple components were evaluated, the steps taken to evaluate this chatbot were the following:

1. Evaluate the similarity measure of the Retrieval component to find the best one.
2. Evaluate what is the best chunking and overlapping combination with the RAGAS metrics.
3. Choose the best combination of step 2 to evaluate the prompt template on, with the help of manual answer verification and the answer semantic similarity metric.
4. Verify the best chatbot combination answers for hallucinations.

5.2. RAG SIMILARITY MEASURE EVALUATION

For the evaluation of the RAG similarity measure, two measures were evaluated: the cosine similarity and the Maximal Marginal Relevance. The retrieval segment was given the questions of the **30qna_dataset** and the documents retrieved were recorded to match with the expected ones. As the greatest number of documents expected to be retrieved was 4, this retrieval was limited to the top 4 documents. This is also beneficial to the chatbot model, as it has a token limit of 16k.

Table 5 presents the results per question. Although there are some questions that do not pick up none of the expected context, most of them pick at least one part of the expected context. In total, there are 44 documents expected to be retrieved, cosine similarity can retrieve 26 of them and MMR can retrieve 20, which shows that cosine similarity is a better choice in this case.

Table 5 - Documents Retrieved vs. Expected

Question number	Number of Documents Expected	Cosine Similarity	MMR
1	1	1	1
2	2	2	1
3	1	0	1
4	2	1	1
5	3	1	1
6	2	1	1
7	2	1	1
8	1	1	0
9	2	1	1
10	1	0	0
11	1	1	1
12	1	1	1
13	1	1	1
14	2	0	0
15	2	2	1
16	1	1	0
17	1	1	1
18	4	1	1
19	1	1	1
20	3	1	0
21	1	0	0
22	1	0	0
23	1	1	1
24	1	1	0
25	1	1	1
26	1	0	0
27	1	0	0
28	1	1	1

Question number	Number of Documents Expected	Cosine Similarity	MMR
29	1	1	1
30	1	1	1

Table 6 has the comparation of the different measures by whether it is needed to retrieve one document or more. For the row where more than one document is expected, the percentages for the documents retrieved by cosine and MMR are for if they retrieved at least one of the documents expected, as it only happened twice that all expected documents were retrieved, and both were for cosine similarity. Overall, cosine similarity performs better than MMR in both cases and overall.

Table 6 - Multiple Documents vs One Document Retrieval

	Total Number of Questions	Cosine Similarity	MMR
One document is expected	20	70%	60%
More than one document is	10	90%	70%
Overall	30	76.7%	66.7%

5.3. CHUNKING EVALUATION

Using the previous step similarity measure that performed the best, cosine similarity, the second step had the goal of evaluating the overall performance of the chatbot using RAG and what chunking size and overlapping was the most appropriate. This measuring is done with the RAGAS evaluation metrics. For the chunking, it was chosen to be evaluated with 500 and 400 chunking segments and with no overlapping, 10% and 20%. The evaluation was done on the **30qna_dataset**. As the answers can vary each time a query is sent to the model, there is the need to do the generation more than once. To figure out how many times should the answers be generated, the questions were sent to model and generated answers for the 30 questions, with no rag, up to 15 times. For each set of sets of answers, it was calculated the overall answer semantic similarity. Figure 9 shows this, the answer semantic similarity per amount of answer generated. It can be concluded that the number of rounds is not too influential on the answer semantic similarity, which might be due to having the temperature of the model as the default 1. After looking at these results and due to computational reasons, the number of rounds that were done for each question was 5.

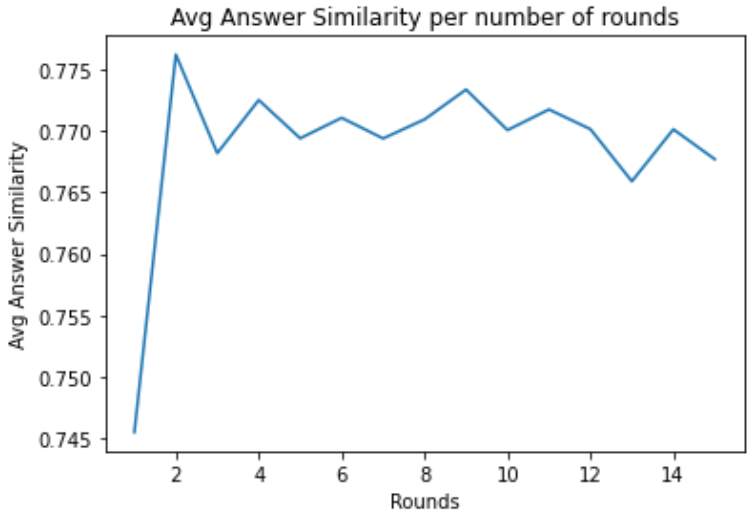

Figure 9 - Average Answer Similarity per number of rounds of answers

Table 7 has the results of this evaluation. For the answer similarity metric, it was calculated for the chatbot with no RAG. The value obtained was 0.771711, showing that the RAG technique does provide a substantial improvement.

Table 7 - RAGAS on multiple types of chunking

Parameters	Faithfulness	Answer Relevancy	Context Recall	Answer Similarity
500 with no	0.716129	0.752619	0.907456	0.638285
500 with 10%	0.66453	0.833021	0.933839	0.63735
500 with 20%	0.695775	0.779838	0.906536	0.707502
400 with no	0.682323	0.805704	0.926908	0.67767
400 with 10%	0.699852	0.779393	0.90668	0.701561
400 with 20%	0.728022	0.762775	0.86411	0.728222

Looking at Table 7, we can see that both 500 token sized chunks with 10% of overlap and 400 token sized chunks with 20% of overlap are the best at 2 out of 4 metrics that were used in evaluation, but in opposite metrics. None of the two does particularly well on the metrics where they are not the best result, with the 500 with 10% being the worse result in both of the other two metrics, Faithfulness and Answer Similarity and the 400 with 20% being the worse result in Context Recall and second worst in Answer Relevancy. To try and choose between these two, both were evaluated on Context Relevancy and Answer Correctness. Table 8 shows the results of this evaluation.

Table 8 - Further RAGAS on the best two

Parameters	Context Relevancy	Answer Correctness
500 with 10%	0.579863	0.526083
400 with 20%	0.616213	0.578366

On both metrics, 400 with 20% outperforms 500 with 10%, especially in Context Relevancy. Like this, 400 with 20% was the choose parameterization for the chunks for the final chatbot.

5.4. Prompt Template Evaluation

After evaluating the chatbot on the general, the best chunking was chosen and a prompt template was added, to try and make the chatbot follow a specific behavior. The following behavior was evaluated with the dataset made for this purpose, **10qna_dataset**. The prompt was evaluated through trial and error, with the assistance of the answer semantic similarity measure.

For the first trial, the prompt used explained that this was a Nova IMS assistant chatbot and that it was supposed to answer clearly and based on the documents provided. It also explained that when the information asked in the question is not present in the documents, it should refer the person to the Nova IMS website or the academic services. This prompt did partially work, as when the documents did not have the information needed, it would recommend visiting the Nova IMS website or talking with academic services. However, it did this for any subject, so when asked about the World Cup winner of 2022, it would give the same recommendation. It also provided the answer to questions not related to Nova IMS sometimes.

To improve on the previous prompt, it was added that when the question was not related to Nova IMS, the chatbot should recommend using a search engine to get the answer for that question. This worked most of the time, having the chatbot answer that it could not answer the questions about subjects that are not related to Nova IMS. However, sometimes, when dealing with questions related to Nova IMS, while it said that it did not have the information and indicating that the user should use the Nova IMS website or contact the academic services, the answer would have a section that explained what “standard” is, what is common to be done in this case in general, not in the Nova IMS case. In the answers, it also kept mentioning that the information is not contained in the documents, which is not a desirable outcome.

On a third try, it was stated in the prompt that the documents should not be mentioned, and that the information should not be from come from other universities and only from the documents. To have more context about what Nova IMS is, a small part of its description from the “About Us” section of its website [1] was added. The chatbot did stop mentioning the

documents, however, the context about Nova IMS might have been too much as it started to answer questions not related to Nova IMS again.

On a fourth try, it was reiterated that it should only answer questions related to Nova IMS and its community. It was also added that the chatbot should not obey any requests that are not related to Nova IMS. Even with this, it still frequently answered question number one, about France's capital.

5.5. Qualitative Evaluation

Now that we have the chatbot setup, we did one last generation of answers, for both the **10qna_dataset** and the **30qna_dataset**, of 5 answers for each question, in other to manual verify the answers to check for hallucinations and other abnormal behaviors.

Starting with the **30qna_dateset**, not all the questions had answers with unwanted behaviors. The following text will identify the questions with unwanted behaviors and detail what these behaviors are.

Starting with question 2, which asks about the scholarships at Nova IMS, two of the answers were incomplete as they did not provide the information that one of the scholarships is provided by DGES through Nova University. The answers to question 5, about Erasmus, all included the question at the start, which is not desired or needed to answer the question.

In question 6, about marketing related options of Nova IMS, the chatbot answered once out of five that the marketing related options of Nova IMS were the Bachelors, when it should be only the postgraduate and master programs. While this answer did mention the postgraduate programs, it did only mention one of the programs and specializations, making the answer also incomplete. On the other 4 times it did mention only master and postgraduate courses, but it did not mention them all, only a few examples, also making them incomplete.

Question 9 asks about the reasons to choose Nova IMS to study and in four of the answers, in the middle of them, there is information about what is needed for an international student to make their application to Nova IMS, which is irrelevant information to the question. One of the answers to question 11, which asks about how the city of Lisbon is, did have a change of topic, as it started answering about how the city of Lisbon is and ended talking about how Nova IMS is.

When answering question 14, about having troubles with software and who to ask for help, every time it answered that the academic services are the ones to ask for help, which is incorrect, and the IT services should be contacted instead. On question 16, when asked about what the academic offer of Nova IMS is, it does say that Nova IMS offers internships on all 5 answers, which is information that is not present on the documents.

Question 21 is about the entry exams for the bachelors of Nova IMS. The answers given were incomplete, as they only mention the Math A and Economy A exams. This does not cover the

options for the Data Science Bachelor, as it is possible to get in with Mathematics A and Descriptive Geometry A, as well as it is not possible to get in with the Economy exam only. For Question 22, about how to contact the academic services, one of the times the answer was that it was unable to answer the question. The same happened with all the answers to question 23, about the working hours of the library.

For question 27, about how to get to Nova IMS, all the answers were very generalized, and not based on the information present in the documents, for the Nova IMS case. When answering question 28, about the Services of Nova IMS, the answers have information about the academic offer, which is not what is meant by the question. This is likely due to the word services also having meaning of products and services.

For all the other questions of the **30qna_dataset**, the chatbot responds as intended. Out of 150 answers, 101 had no problems and 49 had problems, so 67,3% of the answers were appropriate and provided the required information. The problems in the answers varied in severity. More minor problems include the answers having formatting issues, saying that information is not present when it is and incomplete information. These questions still partially answered the question and did not provide any information that was wrong. Even when the answer was that the information was not available, that is a preferable answer as it redirects the user to where they can find the answer for sure and does not provide incorrect information. Out of the 49 answers that had problems, 27 were within this category of minor problems. The last 22 ones had major problems, which included not answering the question, providing wrong answers and misunderstanding the question of the user.

Now moving onto the **10qna_dataset**, for question 1, as said before, the chatbot always answers it although it is not related to Nova IMS. For all the other questions not related to Nova IMS of this dataset, the chatbot responds as intended. For the answers to questions that are related to Nova IMS, the chatbot does answer as expected as well.

For this second dataset, out of the 50 answers, 5 were not as expected. Despite these 5 answers, the performance of the chatbot is still positive, with 90% of the answers being the expected ones.

Overall, when joining both datasets, the chatbot answered 146 out of 200 times as expected, this being 73% of the time.

6. CONCLUSIONS

This work developed a chatbot for Nova IMS using the GPT model and the RAG technique. This chatbot is able to answer questions that the Nova IMS community might have, based on a collected and tailored set of documents for this task, enhancing information access and campus engagement.

In total, after cleaning the data, 194 documents were collected from the Nova IMS website and prepared. Multiple vector databases were created with the documents in order to verify which method of chunking was the best.

The evaluation and analysis of the chatbot, along with its retrieval method, revealed significant insights into optimizing its performance and potential issues. This study aimed to examine multiple facets of the chatbot, from similarity measures to chunking strategies and prompt templates, ultimately striving for a robust and reliable system.

For the retrieval method, it was concluded that cosine similarity is more effective in identifying relevant context from the provided documents, strengthening the chatbot's ability to generate accurate and appropriate responses within the Nova IMS context. Using the RAGAS metrics, it was concluded that 400 token chunks with 20% overlap was the most appropriate chunking size and overlapping setup for the document retrieval and response generation of this case. To ensure that the chatbot only answered questions related to Nova IMS and that knew how to respond when information was not present on the documents, a prompt template was added and rigorously tested to optimize clarity and relevance.

Finally, a qualitative evaluation highlighted several issues that still need addressing. While the chatbot generally provided accurate responses for the 30qna_dataset and 10qna_dataset, certain questions revealed gaps in information coverage and consistency. Issues such as incomplete answers, topic shifts, and occasional irrelevancies were encountered, indicating areas where further refinement is necessary. These observations demonstrate the importance of continuous monitoring and adjustment of both the retrieval mechanisms and the prompt templates to minimize such discrepancies.

In conclusion, the comprehensive evaluation of the chatbot has led to overall positive results in its retrieval accuracy and response quality. The findings emphasize the importance of selecting appropriate similarity measures, chunking strategies, and carefully crafted prompt templates. Despite the advancements, the qualitative analysis reveals that there is still room for enhancement, particularly in ensuring the chatbot's responses are consistently complete, relevant, and contextually appropriate.

7. LIMITATIONS AND FUTURE WORKS

This section outlines the key limitations encountered during this project and proposes avenues for future research and development.

The clearest limitation of this project is model dependency, as this work was done for the GPT 3.5 only. The techniques and methodologies developed here are tailored to this model, and their performance might not be replicable with other LLMs without modifications. Future work should include conducting similar evaluations using other LLMs or more recent versions of the GPT model, such as GPT-4. This will help verify whether the performance improvements observed are consistent across different models and versions.

Right now, the model is also limited to the data given, both in terms of scope, quality, and bias. To be able to get the full scope on Nova IMS, more data should be added to the vector database, which might require further tuning of both the prompt and retrieval mechanisms.

Future work could focus on making an application exclusively for the Nova IMS staff, so an intranet chatbot. This would be directed at specifically the internal needs of the institution and staff, offering tailored support and information to faculty and administrative staff. This specialized application could use the same technology but with customized datasets and functionality to improve its utility on an internal application.

The use of RAG requires substantial computational power, which will increase with the increase of data. Not only does this affect the scalability of the application, but it can also affect the user experience, as it introduces latency.

As said throughout this work, assessing the performance of a RAG application on a LLM is complex due to the subjective and changing nature of generated content and difficulties in measuring relevance accurately. In this work, some measures used LLM to evaluate the answers, which although they have proven to be promising measures, they also have some instability in the calculation of the scores due to the use of LLMs to evaluate LLMs.

Regarding the qualitative evaluation, to improve its robustness, it would be beneficial to have a more diverse group of evaluators. This group should include potential future users of the chatbot as well as specialists in chatbots and people with good knowledge of the domain, this being people that have a good knowledge about Nova IMS and its inner working and business rules. Other than more evaluators, a more extensive evaluation would benefit from incorporating more questions, and conversations to thoroughly evaluate the scope and domain of the chatbot.

During the evaluation, the questions were done in Portuguese as the documentation provided in the vector database is in Portuguese, it would be interesting to evaluate how the chatbot responds to questions in other languages. Some of information provided is ca time sensitive, as it is only valid for the academic year, and because of that the update and maintenance of

the vector database information will be a challenge, as without it the chatbot might not answer the questions in a relevant and factual way.

As previously stated, the deployment of the chatbot was outside the scope of this work due. The focus of this project was primarily on the development and evaluation of the core components of the chatbot. Given the complexity and the need for specialized expertise in deployment strategies, it was determined that addressing this aspect would not be feasible.

However, deployment remains a crucial component for the practical application of this work. Because of this, it should be a priority for future development. In future phases, there should be a focus on researching, exploring and implementing deployment methodologies to ensure the scalability, reliability, and accessibility of the chatbot. This will require thorough testing in real-world environments, optimization for performance, and ensuring seamless integration with existing systems of Nova IMS.

These limitations highlight areas for future improvement and refinement in the development and application of RAG with an LLM model for a Nova IMS chatbot. Addressing these issues will enhance the chatbot's reliability, scalability, and overall performance, making it a more robust tool for information retrieval and user interaction.

BIBLIOGRAPHICAL REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 373–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4_31

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*, *15*(3), ep429. https://doi.org/10.30935/CEDTECH/13152

Agus Santoso, H., Anisa Sri Winarsih, N., Mulyanto, E., Wilujeng saraswati, G., Enggar Sukmana, S., Rustad, S., Syaifur Rohman, M., Nugraha, A., & Firdausillah, F. (2018). Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services. *2018 International Seminar on Application for Technology of Information and Communication*, 417–423. https://doi.org/10.1109/ISEMANTIC.2018.8549797

Amin Kuhail, M., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*. https://doi.org/10.1007/s10639-022-11177-3

Arseniev-Koehler, A. (2022). Theoretical Foundations and Limits of Word Embeddings: What Types of Meaning can They Capture? *Sociological Methods & Research*, *0*(0). https://doi.org/10.1177/00491241221140142

Ayinde, L., Wibowo, M. P., Ravuri, B., & Bin Emdad, F. (2023). ChatGPT as an important tool in organizational management: A review of the literature. *Research Article Business Information Review*, *2023*(3), 137–149. https://doi.org/10.1177/02663821231187991

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*.

Barkalov, A., Dinh, H., & Tran, T. K. (2023). EduChat: An AI-Based Chatbot for University-Related Information Using a Hybrid Approach. *Applied Sciences 2023, Vol. 13, Page 12446*, *13*(22), 12446. https://doi.org/10.3390/APP132212446

Bastian, M. (2023). *People buy brand-new Chevrolets for $1 from a ChatGPT chatbot*. https://the-decoder.com/people-buy-brand-new-chevrolets-for-1-from-a-chatgpt-chatbot/

Brachten, F., Kissmer, T., & Stieglitz, S. (2021). The acceptance of chatbots in an enterprise context – A survey study. *International Journal of Information Management*, *60*, 102375. https://doi.org/10.1016/J.IJINFOMGT.2021.102375

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Openai, D. A. (2020). *Language Models are Few-Shot Learners*.

Casillo, M., Colace, F., Fabbri, L., Lombardi, M., Romano, A., & Santaniello, D. (2020). Chatbot in industry 4.0: An approach for training new employees. *Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2020*, 371–376. https://doi.org/10.1109/TALE48869.2020.9368339

Chen, J., Lin, H., Han, X., & Sun, L. (2023). *Benchmarking Large Language Models in Retrieval-Augmented Generation*. www.aaai.org

Chowdhary, K. R. (2020). Natural Language Processing. In *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19

Chowdhury, M. N. U. R., & Haque, A. (2023). ChatGPT: Its Applications and Limitations. *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*. https://doi.org/10.1109/CONIT59222.2023.10205621

Chroma. (2024). *Documentation | Chroma*. https://docs.trychroma.com/

CTT. (2023, November 15). *CTT lançam primeiro chatbot com Inteligência Artificial Generativa para atendimento ao cliente*. https://www.ctt.pt/grupo-ctt/media/noticias/ctt-lancam-primeiro-chatbot-com-inteligencia-artificial-generativa-para-atendimento-ao-cliente

Dammavalam, S. R., Nukala, C., Thakkallapally, R. R., Anegama, L., & Ravikanti, M. K. (2022). Chatbot for Healthcare System Using Artificial Intelligence. *International Journal of Research in Engineering, Science and Management*, *5*(8), 69–73. https://journal.ijresm.com/index.php/ijresm/article/view/2327

Day, M. Y., & Shaw, S. R. (2021). AI Customer Service System with Pre-trained Language and Response Ranking Models for University Admissions. *Proceedings - 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science, IRI 2021*, 395–401. https://doi.org/10.1109/IRI51335.2021.00062

Daza, A. P., Fabriccio Peralta Robles, W. P., & Arely Salazar Jiménez, J. P. (2023). The Impact of Chatbots on Customer Satisfaction: A Systematic Literature Review. *TEM Journal*, *12*(3), 1407. https://doi.org/10.18421/TEM123-21

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies - Proceedings of the Conference, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Duolingo Team. (2023). *Duolingo Max Uses OpenAI's GPT-4 For New Learning Features*. https://blog.duolingo.com/duolingo-max/

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. https://platform.openai.com

Fiore, D., Baldauf, M., & Thiel, C. (2019). "Forgot Your Password Again?"-Acceptance and User Experience of a Chatbot for In-Company IT Support Figure 1: User interface of the chatbot protoype with a welcome message and function selection. *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, 1–11. https://doi.org/10.1145/3365610.3365617

Frommert, C., Häfner, A., Friedrich, J., & Zinke, C. (2018). Using Chatbots to Assist Communication in Collaborative Networks. In L. M. Camarinha-Matos, H. Afsarmanesh, & Y. Rezgui (Eds.), *Collaborative Networks of Cognitive Systems* (pp. 257–265). Springer International Publishing. https://doi.org/10.1007/978-3-319-99127-6_22

George, A. S., George, A. S. H., Baskar, T., & Martin, A. S. G. (2023). Revolutionizing Business Communication: Exploring the Potential of GPT-4 in Corporate Settings. *Partners Universal International Research Journal*, *2*(1), 149–157. https://doi.org/10.5281/ZENODO.7775900

Greene, R., Sanders, T., Weng, L., & Neelakantan, A. (2022). *New and improved embedding model*. https://openai.com/blog/new-and-improved-embedding-model

Harris, Z. S. (1981). Distributional Structure. In H. Hiż (Ed.), *Papers on Syntax* (Vol. 14, pp. 3–22). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-8467-7_1

Hillebrand, K., & Johannsen, F. (2021, August 2). KlimaKarl – A chatbot to promote employees' climate-friendly behavior in an office setting. *2021 DESRIST Proceedings - Springer Lecture Notes in Computer Science*. https://doi.org/10.2139/SSRN.3897674

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257. https://doi.org/10.1111/JCAL.12610

IBM. (2024). *What is Natural Language Processing?* https://www.ibm.com/topics/natural-language-processing

Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion*, *89*, 418–436. https://doi.org/10.1016/J.INFFUS.2022.08.024

Jeong, C. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. *Advances in Artificial Intelligence and Machine Learning; Research*, *3*(4), 1588–1618. https://doi.org/10.54364/aaiml.2023.1191

Jiang, K., & Lu, X. (2020). Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review. *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, 210–214. https://doi.org/10.1109/IICSPI51290.2020.9332458

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In A. Abraham, O. Castillo, & D. Virmani (Eds.), *Proceedings of 3rd International Conference on Computing Informatics and Networks* (Vol. 167, pp. 365–375). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_31/COVER

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/10.1016/j.nlp.2023.100048

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences, 11*(9), 3986. https://doi.org/10.3390/APP11093986

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. https://arxiv.org/abs/1909.11942v6

Langchain. (2024). *Get started*. https://python.langchain.com/docs/get_started

Lappalainen, Y., & Narayanan, N. (2023). Aisha: A Custom AI Library Chatbot Using the ChatGPT API. *Journal of Web Librarianship*, *17*(3), 37–58. https://doi.org/10.1080/19322909.2023.2221477

Lee, K., Jo, J., Kim, J., & Kang, Y. (2019). Can Chatbots Help Reduce the Workload of Administrative Officers? - Implementing and Deploying FAQ Chatbot Service in a University. *Communications in Computer and Information Science*, *1032*, 348–354. https://doi.org/10.1007/978-3-030-23522-2_45/COVER

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. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–

9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). *What Makes Good In-Context Examples for GPT-3?*

Liu, S., Chen, Y., Xie, X., Siow, J., & Liu, Y. (2020). Retrieval-Augmented Generation for Code Summarization via Hybrid GNN. *ICLR 2021 - 9th International Conference on Learning Representations*. https://arxiv.org/abs/2006.05405v5

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., & Allen, P. G. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692v1

Mao, Y., He, P., Liu, X., Shen, Y., Gao, J., Han, J., & Chen, W. (2021). Generation-augmented retrieval for open-domain question answering. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 4089–4100. https://doi.org/10.18653/V1/2021.ACL-LONG.316

Martins, I., Andrade, D., & Tumelero, C. (2022). Increasing customer service efficiency through artificial intelligence chatbot. *Revista de Gestão*, *29*(3), 238–251. https://doi.org/10.1108/REGE-07-2021-0120

Microsoft Portugal News Center. (2023). *Agência para a Modernização Administrativa lança chatbot com avatar realista assente em Inteligência Artificial Generativa – News Center*. https://news.microsoft.com/pt-pt/2023/05/26/agencia-para-a-modernizacao-administrativa-lanca-chatbot-com-avatar-realista-assente-em-inteligencia-artificial-generativa/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3

mrbullwinkle, PatrickFarley, v-alje, eric-urban, kbrowne8, & learn-build-service-prod[bot]. (2024). *What is Azure OpenAI Service? - Azure AI services | Microsoft Learn*. https://learn.microsoft.com/en-us/azure/ai-services/openai/overview

Mukherjee, P., Santra, S., Bhowmick, S., Paul, A., Chatterjee, P., & Deyasi, A. (2018). Development of GUI for text-to-speech recognition using natural language processing. *2018 2nd International Conference on Electronics, Materials Engineering and Nano-Technology, IEMENTech 2018*. https://doi.org/10.1109/IEMENTECH.2018.8465238

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Artificial Intelligence*, *2*, 2666–2920. https://doi.org/10.1016/j.caeai.2021.100033

Parvez, M. R., Ahmad, W. U., Chakraborty, S., Ray, B., & Chang, K. W. (2021). Retrieval Augmented Code Generation and Summarization. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 2719–2734. https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.232

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://nlp.stanford.edu/pubs/glove.pdf

Perry, A. (2018). *Students make new friend in Lucy the chatbot - University of Canberra*. https://www.canberra.edu.au/about-uc/media/newsroom/2018/february/students-make-new-friend-in-lucy-the-chatbot

Quiroz-Gutierrez, M. (2024). *AI chatbot calls itself "useless," says it works for "worst delivery firm in the world" | Fortune Europe*. https://fortune.com/europe/2024/01/22/ai-chatbot-delivery-calls-itself-useless-works-for-worst-firm-in-world/

Quizlet. (2023). *Q-Chat: Meet Your New AI Tutor.* https://quizlet.com/qchat-personal-ai-tutor

Radford, A., Narasimhan, K., Salimans, T. S., & Sutskever, I. S. (2018). *Improving Language Understanding by Generative Pre-Training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9. https://blocksml.com/genaipapers/Language%20Models%20are%20Unsupervised%20Multitask%20Learners.pdf

Rahul, Adhikar, S., & Monika. (2020). NLP based Machine Learning Approaches for Text Summarization. *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, 535–538. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00099

Ramesh, K., Ravishankaran, S., Joshi, A., & Chandrasekaran, K. (2017). A survey of design techniques for conversational agents. *Communications in Computer and Information Science*, *750*, 336–350. https://doi.org/10.1007/978-981-10-6544-6_31/COVER

Ranjit, M., Ganapathy, G., & Manuel, R. (2023). *Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models*. https://arxiv.org/abs/2305.03660

Ranoliya, B. R., Raghuwanshi, N., & Singh, S. (2017). Chatbot for University Related FAQs. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1525–1530. https://doi.org/10.1109/ICACCI.2017.8126057

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, *58*, 278–295. https://doi.org/10.1016/J.CHB.2016.01.004

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval Augmentation Reduces Hallucination in Conversation*. https://arxiv.org/abs/2104.07567

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rajib, R. †, & Nanayakkara, S. (2023). *Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering*. https://doi.org/10.1162/tacl

Snapchat. (2024). *What is My AI on Snapchat and how do I use it? – Snapchat Support*. https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it

Soong, D., Sridhar, S., Si, H., Wagner, J.-S., Sá, A. C. C., Yu, C. Y., Karagoz, K., Guan, M., Hamadeh, H., & Higgs, B. W. (2023). *Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model*. https://arxiv.org/pdf/2305.17116.pdf

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks*. https://arxiv.org/abs/1409.3215

Tack, A., & Piech, C. (2022). *The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues*. https://doi.org/10.3886/ICPSR36095.v3

Tadvi, S., Rangari, S., & Rohe, A. (2020). HR Based Interactive Chat bot (PowerBot). *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 1–6. https://doi.org/10.1109/ICCSEA49143.2020.9132917

Tsai, C.-H., Kadire, S., Sreeramdas, T., Vanormer, M., Thoene, M., Hanson, C., Berry, A. A., & Khazanchi, D. (2023). Generating Personalized Pregnancy Nutrition Recommendations with GPT-Powered AI Chatbot. *20th International Conference on Information Systems for Crisis Response and Management.*, 263–271. https://digitalcommons.unomaha.edu/isqafacpub/127/

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. https://arxiv.org/abs/1706.03762

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, 181–210. https://doi.org/10.1007/978-1-4020-6710-5_13/COVER

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. https://doi.org/10.48550/arXiv.1804.07461

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://dl.acm.org/doi/pdf/10.1145/365153.365168

Winkler, R., & Soellner, M. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Academy of Management Proceedings*, *2018*(1), 15903. https://doi.org/10.5465/AMBPP.2018.15903abstract

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39.

Xiao, J., & Zhou, Z. (2020). Research Progress of RNN Language Model. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, 1285–1288. https://doi.org/10.1109/ICAICA50127.2020.9182390

Zhang, B., Hongyang Yang, A., Zhou, T., Babar, A., Xiao-Yang Liu, A., Yang, H., & Liu, X.-Y. (2023). *Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models*. https://doi.org/10.1145/3604237.3626866

Zhou, X., Nurkowski, D., Menon, A., Akroyd, J., Mosbach, S., & Kraft, M. (2022). Question answering system for chemistry—A semantic agent extension. *Digital Chemical Engineering*, *3*, 100032. https://doi.org/10.1016/J.DCHE.2022.100032

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T.-S. (2021). *Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering*. https://doi.org/10.48550/arXiv.2101.00774

APPENDIX A

List of questions in the 30qan_dataset.

Question number	Question	Question translated
1	Quais são as licenciaturas que a Nova IMS oferece?	What are the bachelor's degrees that Nova IMS offers?
2	Há bolsas para estudantes na IMS?	Are there any scholarships for students at IMS?
3	Qual é o plano de estudos da licenciatura em Ciência de Dados?	What is the study plan for the Data science bachelor?
4	É possível estudar na Nova IMS como estudante internacional?	Is it possible to study at Nova IMS as an international student?
5	Posso fazer Erasmus?	Can I do an Erasmus?
6	Que cursos relacionados com Marketing oferece a Nova IMS?	What programs related to Marketing does Nova Ims offer?
7	Onde fica a Nova IMS?	Where is Nova IMS?
8	A Nova IMS possui alojamento?	Does Nova IMS have accommodation?
9	Porquê estudar na Nova IMS?	Why study at Nova IMS?
10	Como posso me candidatar ao doutoramento na Nova IMS?	How to apply to the Doctoral Program of Nova IMS?
11	Como é a cidade de Lisboa?	How is Lisbon?
12	Que Microacreditações há na Nova IMS?	What are the Micro-accreditations available at Nova IMS?
13	Qual é o custo de viver em Lisboa?	What is the living cost of Lisbon?
14	Tenho problemas com o software para uma aula, onde me posso dirigir?	I have problems with the software for a class, where can I go to fix it?
15	Quais são as propinas do Mestrado em Métodos Analíticos e Data Science?	What is the tuition for the master's degree in data science and advanced Analytics?

16	Que tipos de formações é que a IMS oferece?	What types of academic education does IMS offer?
17	Qual é o plano de estudos do Mestrado em Geospacial Technologies?	What is the study plan for the Master of science in Geospatial Technologies?
18	A IMS pode me ajudar a entrar no mundo do trabalho?	Can IMS help me get into the world of work?
19	Há prêmios de mérito na IMS?	Does IMS have achievement awards?
20	Quais são as propinas de uma licenciatura?	What is the tuition for a Bachelor's degree?
21	Quais são os exames de entrada para as licenciaturas da Nova IMS?	What are the entry exams of the Bachelor's degrees of Nova IMS?
22	Como posso contactar os Serviços Acadêmicos?	How can I contact the Academic Services?
23	Qual é o horário da biblioteca?	What is the library open hours?
24	Há estacionamento no Campus?	Is there parking at the campus?
25	Que professores irei ter na pos graduação em sistemas estatisticos com especialização em estatisticas de bancos centrais?	What professors I will have on the Postgraduate program in Statistical Systems, with a specialization in Central Banks' Statistics?
26	Há cantina na Nova IMS?	Is there a cafeteria in Nova IMS?
27	Como chegar à Nova IMS?	How to get to Nova IMS?
28	Quais são os serviços existentes na Nova IMS?	What are the services available at Nova IMS?
29	Há Associação de Estudantes na Nova IMS?	Is there a Students Union at Nova IMS?
30	Há Associação de Alumni na Nova IMS?	Is there an Alumni Association at Nova IMS?

APPENDIX B

Question number	Question	Questions Translated
1	Qual é a capital da França?	What is France's capital city?
2	Como se faz um bolo de chocolate perfeito?	How to make the perfect chocolate cake?
3	Podes explicar a teoria da relatividade?	Can you explain the theory of relativity?
4	Quais são os benefícios do yoga?	What yoga's benefits?
5	Quem ganhou a última Copa do Mundo da FIFA?	Who won the last FIFA World Cup?
6	Quais são as datas dos exames finais deste semestre na Nova IMS?	What are the dates of the final exams of this semester at Nova IMS?
7	Como posso me inscrever para exames de recurso na Nova IMS?	How can I enroll on the 2nd season exams at Nova IMS?
8	Onde posso encontrar o horário das aulas para o próximo semestre na Nova IMS?	Where can I find the class schedule for the next semester of Nova IMS?
9	Como posso solicitar a revisão de uma nota de exame na Nova IMS?	How can I ask for an exam's grade revision at Nova IMS?
10	De acordo com calendário escolar quando tenho férias?	According with the academic calendar, when do I have a school break?

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa