# Hello, Freire! How Can You Help Me? A Smart Chatbot to Enhance Access to Academic Opportunities and Institutional Matters

Arthur Willame Mesquita
Federal University of Ceará (UFC)
Itapajé, CE, Brazil
arthurwillame@alu.ufc.br

Carlos Freire
Federal University of Ceará (UFC)
Itapajé, CE, Brazil
carlosjefte@alu.ufc.br

Antônio Gomes
Federal University of Ceará (UFC)
Itapajé, CE, Brazil
antoniocruz@alu.ufc.br

Bruna Amazonas
Federal University of Ceará (UFC)
Itapajé, CE, Brazil
bruna.amazonas@ufc.br

Anderson Uchôa
Federal University of Ceará (UFC)
Itapajé, CE, Brazil
andersonuchoa@ufc.br

## Abstract

**Context:** Universities offer various resources and services that students often fail to utilize due to difficulties in accessing relevant information. Chatbots can improve access and interaction between students and institutions. **Problem:** Students struggle to find information about academic opportunities and institutional matters due to fragmented data, limiting engagement. **Solution:** This paper introduces *Freire Assistant* (Freire), a chatbot system powered by a Large Language Model (LLM) and built using Retrieval Augmented Generation (RAG) pipelines. Freire aims to streamline access to academic information through a fast, intuitive communication channel. **IS theory:** This work was developed under the aegis of Soft Systems Theory, as it addresses the complexity of accessing academic information as a sociotechnical problem. **Method:** To evaluate Freire's effectiveness in providing accurate and relevant information about academic opportunities and institutional matters, we conducted a case study at a Brazilian university. We quantitatively evaluated the contextual quality of Freire's responses using the DeepEval framework. Additionally, we qualitatively evaluated Freire's usability with 45 students through the Chatbot Usability Questionnaire (CUQ) which measures aspects related to Chatbot Personality, Onboarding, User Experience, and Error Handling. **Summary of Results:** The findings demonstrated Freire's efficiency in generating contextually relevant and accurate responses. Furthermore, usability scores were generally high due to their simple interface and conversation-driven functionality. **Contributions and Impact in the IS area:** This study highlights how smart chatbots can enhance user experience and access to information in universities, providing a replicable model for other institutions facing similar challenges.

## CCS Concepts

• **Information systems** → **Information systems applications**;
• **Computing methodologies** → **Natural language generation**.

## Keywords

chatbot, education context, innovation, students, llm

## 1 Introduction

Universities play a critical role in society, supporting teaching, research, and extension while engaging with external communities. However, Brazilian universities face significant challenges due to decentralized structures, leading to fragmented information systems that hinder access to information for students and external communities. These include essential information such as events, scholarship opportunities, and updates on new technologies [18, 31].

To address these challenges, intent-based chatbots have traditionally been used to improve communication, but they struggle with handling complex, open-ended queries and scaling to diverse user needs [6, 29]. In the educational context, such systems often do not provide context-sensitive responses, making it difficult for students to navigate academic programs, and institutional processes effectively. In contrast, LLM-based chatbots, enhanced with Retrieval-Augmented Generation (RAG), offer context-sensitive responses by retrieving and integrating relevant information [14, 22].

In this context, based on the potential of LLM-based chatbots, this paper introduces the Freire Assistant, an advanced chatbot system powered by LLM (LLaMA-3.2-90B) and RAG pipelines, developed through a case study at a Brazilian university. The Freire Assistant was designed to assist students in finding information about academic opportunities and institutional matters. The Freire Assistant provides interactive, real-time responses to queries about professors, the academic calendar, enrollment and procedures, and scholarship announcements. To evaluate its effectiveness, we conducted a comprehensive evaluation involving both quantitative (using a set of well-known metrics [11]) and qualitative (using the Chatbot Usability Questionnaire (CUQ) [10]) methods.

In summary, the main contributions of this paper include: (1) the identification and analysis of communication challenges in decentralized university environments, validated through a structured survey across diverse user groups; (2) the presentation of a scalable framework that can be adapted to other educational institutions facing similar challenges, providing a blueprint for integrating AI-driven solutions in university settings; and (3) the demonstration of the chatbot's ability to improve access to academic information while enhancing the student experience at a Brazilian university.

## 2 Background and Related Work

### 2.1 Large Language Models (LLMs)

LLMs are characterized by their capacity to generate human-like text using vast numbers of parameters, often exceeding hundreds of billions [3]. Examples of LLM models include GPT-4 [26], and Llama3 [31]. These models are typically pre-trained on large datasets, allowing them to learn human language's complexities through unsupervised learning [3]. After pre-training, the models can be

fine-tuned on specific tasks, making them highly adaptable for various domain applications including education. Their large parameter size enables them to capture intricate language patterns, improving their ability to perform complex tasks such as answering questions, translating languages, and summarizing content [29].

Although pre-trained LLMs are proficient at acquiring extensive knowledge, they struggle with the lack of memory expansion due to the temporal limitations of their knowledge base, leading to errors such as hallucinations [34]. To address these challenges, LLMs provide a solution based on the use of RAG [15]. RAG is a technique that efficiently extracts relevant information from unstructured documents when responding to user queries [15]. In this context, RAG-based systems dynamically retrieve information from a large corpus of documents to generate customized responses tailored to the user's specific requests [15]. As a result, the responses are more accurate and contextually enriched compared to intent-based systems that rely on predefined intents and responses, limiting their flexibility and contextual comprehension [19].

More specifically, RAG employs integrated retrieval and generation techniques that collaboratively enhance contextual understanding. The retrieval component identifies the top k most relevant text passages for a given input query, improving both the model's comprehension and useful response generation. This process is represented by the equation $p_n(z|x)$, where $p_n$ represents the retrieval component, $n$ is the number of documents or passages to be retrieved, and $z$ represents the relevant passages selected from the vector databases based on the input $x$ [24].

## 2.2 Chatbot Systems in Educational Settings

Chatbot systems are beneficial in educational settings, where the diversity and complexity of student queries often exceed the capabilities of intent-based chatbots. These intent-based chatbots are typically designed to identify and analyze user intentions using predefined patterns and commands [19]. In contrast, LLM-based chatbots can generate human-like responses, enabling them to handle a non-predefined set of questions in more dynamic and open-ended interactions with students [6]. In this context, we overview some solutions of chatbot systems in educational settings as follows.

**Intent-based chatbots.** Chien and Yao [7] developed a userbot system integrating intent-based and flow-based dialogue modes, enabling engineering students to engage in participatory design activities by interacting with virtual users. Sophia and Jacob [30] introduced Edubot, a chatbot designed to assist students during the COVID-19 pandemic. It aimed to address academic queries through predefined intents, facilitating learning through a question-and-answer format. Similarly, Assayed et al. [1] developed the HSchatbot, a chatbot that supports high school students by predicting the intent of their inquiries related to academic decisions, such as scholarships, university requirements, majors, and curricula. Additionally, Wang et al. [32] introduced an educational chatbot that combines intent classification and slot-filling models to enhance online learning experiences by understanding task-oriented natural language texts to provide education-related services.

**LLM-based chatbots.** Neupane et al. [24] introduced BARKPLUG V.2, a chatbot powered by LLM and RAG pipelines, developed at Mississippi State University. The BARKPLUG V.2 aimed to generate accurate and relevant responses to specific questions about the university such as academic departments, programs, campus facilities, and student resources. Using RAG, BARKPLUG V.2 has significantly enhanced user satisfaction and engagement [24]. Similarly, Chandra and Suyanto [5] developed a chatbot to guide the admissions process at an Indonesian university. Oliveira and Matos [25] introduced a dynamic chatbot designed to enhance student interaction by addressing inquiries on admissions, academic support, and event information. The chatbot was built to prioritize user feedback for accuracy, reliability, and safety.

Liu et al. [18] integrated a RAG-based solution into an introductory course in computer science at Harvard University. The authors observed the effectiveness of RAG in providing detailed and contextually appropriate responses, enabling personalized tutoring for students. Additionally, Martinez-Araneda et al. [21] introduced TutorBot+, aimed at delivering feedback in programming courses. TutorBot+ demonstrated positive impacts on students' computational reasoning abilities, demonstrating the potential of AI-driven interventions in education. In summary, chatbots in educational settings facilitate academic processes, promptly support students, and enhance academic experiences.

In a nutshell, our work differs from the existing ones in the following points: (i) we integrate an improved RAG pipeline into the *Freire Assistant* to ensure more precise and contextually relevant responses; (ii) we extend the *Freire Assistant* beyond departmental inquiries and admissions to also support academic opportunities and administrative procedures; and (iii) we conduct a rigorous assessment of the *Freire Assistant*, combining quantitative metrics (DeepEval) and qualitative evaluation (CUQ) to ensure a well-balanced focus on technical performance and user experience.

# 3 Freire Assistant at a Glance

This section overviews the Freire Assistant, including key features (Section 3.1), and architecture/design decisions (Section 3.2).

## 3.1 Freire Features

Figure 1 illustrates the main features of the Freire Assistant, including a dialogue in English[1]. We describe the features as follows.

**Theme Switching.** The Freire offers a customizable interface, allowing users to switch between light and dark themes, as illustrated in item ①. This feature ensures visual comfort across diverse environments and user preferences, aligning with best practices for accessibility in user interface design. **Predefined Query Cards.** To simplify information retrieval, the chatbot provides predefined query cards containing: (i) a short topic title and (ii) a description for each topic. In this context, users can select these cards to receive immediate and detailed responses. For instance, as shown in ②, the cards cover essential topics such as enrollment and procedures, academic calendars, and scholarship opportunities, enabling quick answers to relevant and frequent information.

**Human-AI interaction.** The chatbot leverages advanced natural language capabilities to interpret user inputs in natural language, whether provided as typed text or transcribed speech (items

[1]Since the study was conducted at a Brazilian university, the dialogue was translated into English for presentation purposes.

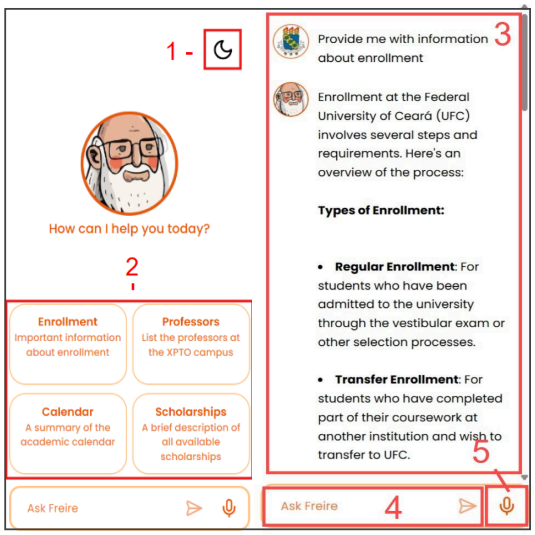


**Figure 1: Overviews of the Freire Assistant Features**

③ and ④). This allows the system to dynamically handle both open-ended and context-specific queries. For instance, as shown in the figure on the right, a user's question about enrollment prompts the chatbot to provide a detailed explanation of the types and procedures for enrollment. **Speech-to-Text Conversion.** The chatbot features speech recognition, enabling users to interact via voice input through their device's microphone (item ⑤). Spoken inputs are transcribed into text using the Web Speech API [9] , allowing the chatbot to process the message and generate a response.

## 3.2 Freire Architecture and Design Decisions

Figure 2 illustrates the main architectural components of the Freire.

**Data Administration, Input, and Extraction.** To consolidate Freire's knowledge base, we implemented a data administration, input, and extraction system that handles unstructured documents. The system is designed to support PDF format only, which requires a process of extraction and conversion to text strings, ensuring that unstructured content can be processed. In our case, users with the roles of *administrator* or *uploader* can upload up to 10 documents simultaneously, which are then converted to text. For text extraction, we used the *pdf-parse*[2] library (for PDFs). The documents may vary in structure and quality, requiring robust preprocessing to handle noise (e.g., unwanted headers or footers). In this context, we removed HTML tags and extra spaces, along with the default preprocessing of Jina.ai [12], which we used to separate the title, URL source, and markdown content.

Additionally, to enhance Freire's knowledge base, we integrated a web scraping mechanism that dynamically retrieves relevant university information when no matching data is found in the vector database. Instead of pre-scraping and storing data in JSON files, our approach leverages real-time web scraping using Jina.ai, which fetches content from predefined university-related links. If a query yields no relevant results from ChromaDB, the system scrapes the corresponding webpage, extracts its content, and immediately adds it to the vector database for future queries. This ensures that Freire Assistant always has access to the latest information, including professor descriptions, the academic calendar, and enrollment procedures. Table 1 provides an overview of the data sources used by Freire Assistant. The entire dataset is in Portuguese.

**Table 1: A subset of data source utilized by Freire**

| Category | # of Tokens |
|---|---|
| Professors | 6.038 |
| Academic Calendar | 10.169 |
| Enrollment and Procedures | 2.982 |
| Scholarship Announcements | 218.671 |
| **Total** | **237.860** |

**Data Segmentation into Chunks.** The source data are processed and divided into smaller units, known as chunks. These chunks typically consist of sentences or paragraphs [14]. We adopted `RecursiveCharacterTextSplitter` from Langchain [13], which ensures adaptable segmentation. This technique processes the text by first dividing it using higher-level delimiters, such as paragraphs or newlines [14]. For over-sized chunks, it recursively applies lower-level delimiters, such as sentences or words, until all chunks meet the desired size constraint. The parameters were set to 1000 for chunk size and 200 for overlap. Additionally, a defined overlap between chunks maintains context across boundaries. This approach creates an array of strings that facilitates the search and retrieval of contextual information within the vector database.

**Data Vectorization (Embeddings).** The text data at the segment level is transformed into numerical vector representations using embeddings. Thus, an embedding model converts each chunk into a vector representation. This step is critical to ensure the system can perform efficient contextual searches, capturing the content's semantics, which is essential for similarity-based retrieval. These embeddings are stored in the vector database, with the API integrated into the system to optimize latency and manage batch calls, thereby reducing costs. We used the `Sentence Transformers all-MiniLM-L6-v2 model` [3] to generate embeddings. This embedding model can create both sentence and document embeddings, applicable to a wide range of tasks. The embedding function runs locally on the machine and may require model file downloads, which happen automatically if needed.

**Vector Database.** We used ChromaDB[4] to store generated vectors. ChromaDB is designed for rapid indexing, enabling efficient retrieval of semantically similar data in response to user queries while ensuring data security and reliability in transformed data storage. Each generated vector is stored alongside an identifier for the corresponding text chunk, allowing similarity searches to retrieve relevant content. The vector index is updated when new documents are added, and hierarchical navigable small-world graph-based indexing [20] is used to optimize query response times with cosine similarity as the distance metric.

**Information and Context Retrieval.** To retrieve contextually relevant information based on the prompt, five specific chunks are

[2]https://gitlab.com/autokent/pdf-parse

[3]https://docs.trychroma.com/guides/embeddings#default:-all-minilm-l6-v2-8

[4]https://www.trychroma.com/

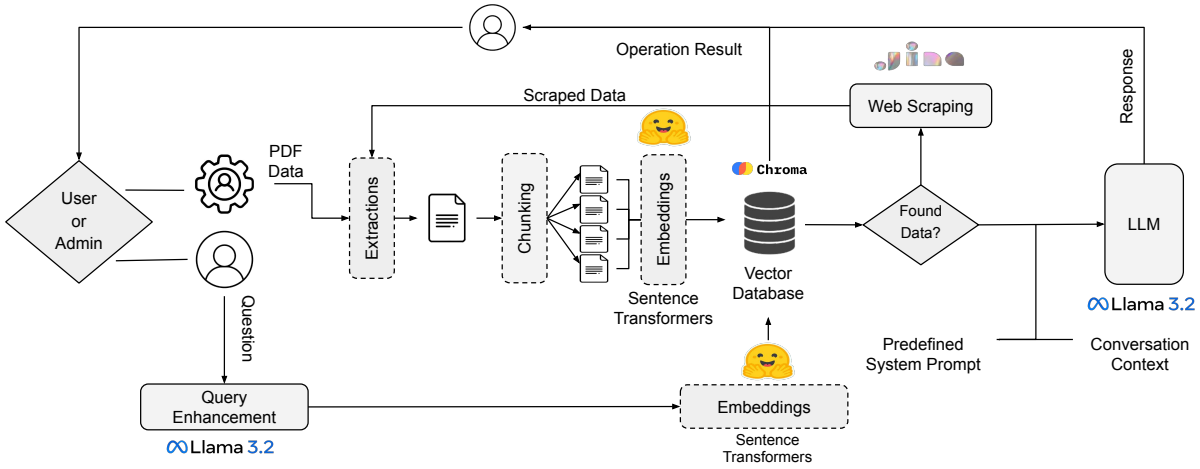


**Figure 2: Freire architecture**

fetched from the vector database. In this context, the user's query is first enhanced using an LLM (LLaMA-3.2-90B via Groq API) to improve semantic search, then processed and transformed into a vector, which is compared with vectors in the vector database to identify the most relevant chunks. We chose LLaMA-3.2-90B due to its popularity, availability, scalability, and cost. The decision to retrieve multiple relevant chunks allows for composing a more robust context for the language model in generating the final response. Additionally, retrieving chunks instead of entire documents increases response precision and reduces information overload, as the LLM receives only the most relevant portions of the content.

**Response Generation by the Language Model.** The final stage is generating responses based on the user's query and the retrieved chunks. In this context, the query and retrieved chunks are integrated into a single prompt, which is sent to the LLM. Parameters such as response type and length can be configured to adapt the response style to the user's needs. The architecture is designed to allow dynamic adjustments to the prompt, depending on the usage context and specific application requirements. Integrating context with the query can result in lengthy prompts, which may pose challenges for models with token limits. To address this, we implemented an intelligent truncation and prioritization strategy, ensuring that the LLM receives the essential context.

**Context Management.** We used LangChain to manage conversation context through the Assistant API. LangChain offers a comprehensive suite of tools for implementing LLMs, among which is the Assistant API. This API encapsulates all messages exchanged between the user, the system, and the LLM, formatting them into a structured chat format. This enables the model to comprehend the conversational context and recall previous messages, ensuring continuity and coherence in interactions.

**Technologies Used.** For the chatbot implementation, we adopted modern technologies that provide robustness, flexibility, and performance. The user interface was developed with React in conjunction with TypeScript, which allowed for the creation of an interactive and intuitive experience, while the static typing provided by TypeScript reduced errors and facilitated code maintenance and evolution. On the backend, Node.js with TypeScript and the Express framework were used, leveraging their scalability and flexibility to process data and integrate the different parts of the system.

## 4 Study Settings

To define the goal, and research questions (RQs), we followed the Goal-Question-Metric (GQM) template [4]. Our goal is: **analyze** the effectiveness of the Freire Assistant chatbot; **for the purpose of** enhancing student access to academic opportunities and institutional matters; **with respect to** (i) the quality and relevance of the responses provided, and (ii) usability; **from the viewpoint of** undergraduate students; **in the context of** an academic unit of a Brazilian university. We detail each RQ as follows:

**(Quantitative evaluation) $RQ_1$:** *To what extent does the Freire Assistant generate contextually accurate and relevant responses about academic opportunities and institutional matters?* – $RQ_1$ aims to explore the chatbot's ability to generate responses that are both accurate and contextually relevant. Given that the Freire Assistant uses a LLM in conjunction with a RAG pipeline, we can measure how effectively the chatbot can retrieve and generate information about academic opportunities and institutional matters (Professors, Academic Calendar, Enrollment and Procedures, and Scholarship Announcements). To answer $RQ_1$, we used the DeepEval framework, designed to quantitatively assess the quality of the responses in terms of their correctness, relevance, and context (see Section 4.1).

**(Usability evaluation) $RQ_2$:** *How do students perceive the usability of the Freire Assistant?* – $RQ_2$ aims to evaluate the chatbot's usability based on the user-side experience of interacting with Freire. To this end, we employed the Chatbot Usability Questionnaire (CUQ) to obtain insights from the participants, related to four aspects: (1) the chatbot's personality, (2) the onboarding process, (3) user interface, and (4) how effectively it handles errors (see Section 4.2). By answering $RQ_2$, we can determine whether

students find the Freire Assistant intuitive and helpful in accessing academic opportunities, contributing to the broader goal of improving student engagement with university resources.

## 4.1 Quantitative Evaluation

To evaluate the contextual quality of Freire's response, we utilized the DeepEval framework [11]. This framework provides metrics for evaluating the precision, recall, and relevancy of generated text in relation to the expected outputs and the context provided by the retrieved documents [11]. We selected DeepEval because it is specifically designed to assess RAG pipelines. In contrast, other frameworks such as ROUGE [17] and BLEU [27] are not suitable in our context, as ROUGE is typically used for summarization tasks, while BLEU is designed to evaluate the language translation tasks.

The evaluation process involves *selecting appropriate metrics* and *creating a set of questions with their ground truth* as test cases. These test cases are used to assess the chosen metrics by comparing the input query, the system's actual output, the expected output (generated using LLaMA-3.2-90B), and the retrieval context. This ensures a structured and accurate evaluation of the system's performance.

**Selection of appropriate metrics.** We selected five key metrics from the DeepEval framework [11] to assess the quality of the generated responses. Table 2 provides an overview of each metric.

**Table 2: Metrics selected to evaluate the contextual quality**

| Name | Description |
|---|---|
| Contextual Precision | Measures how many of the retrieved documents contain information that is relevant to the generated response. |
| Contextual Recall | Measures the proportion of relevant information in the retrieved documents that is used in the generated response. |
| Contextual Relevancy | Assesses how relevant the generated response is to the query and the context provided by the retrieved documents. |
| Faithfulness | Evaluate how well a model-generated response aligns with the retrieved documents and stays true to the source information. |
| Answer Relevancy | Evaluates the degree to which the generated response addresses the user's query based on the retrieved documents, ensuring that the output is both contextually appropriate and factually relevant to the input. |

We employed the *Contextual Precision*, *Contextual Recall*, and *Contextual Relevancy* metrics to evaluate the efficiency of the retrieval process and the contextual relevance of the generated responses. Additionally, we used *Faithfulness* to assess the factual accuracy of the generated responses, while *Answer Relevancy* measured how well the responses addressed the original query. Furthermore, we selected metrics to conduct an end-to-end evaluation of different dimensions of Freire's responses regarding user experience. These metrics assess aspects from coherence and clarity to contextual appropriateness and adherence to user instructions, capturing essential aspects of response quality in human-AI interaction. These metrics were adapted from Deep Eval's end-to-end metrics, focusing on the students. Table 3 overviews of these metrics.

**Creation of a set of questions and their ground truth as test cases.** We developed a set of questions and their corresponding reference answers, serving as a benchmark to evaluate the effectiveness and relevance of generated responses. This ensures that future evaluations are aligned with real user experiences. To elicit this set of questions, we surveyed 31 incoming students, 49 veteran students, and 35 university staff, totaling 115 participants to capture the needs and expectations concerning the context being evaluated.

**Table 3: Metrics used for end-to-end evaluation**

| Name | Description |
|---|---|
| Coherence | Assesses how logically connected and consistent the sentences in the output are, ensuring a smooth and cohesive flow of ideas. |
| Clarity | Evaluates how easy the output is to understand, avoiding complex language and ambiguity. |
| Context Appropriateness | Measures how well the response fits within the specific context of the user's query and previous interactions. |
| Instruction Following | Measures how accurately the output follows the instructions given in the user's query. |

The survey included questions about the communication channel and the difficulties in finding specific information about academic opportunities and institutional matters. After data collection, we conducted a quantitative analysis of the closed-ended responses, and a qualitative analysis of the open-ended responses to categorize emerging common questions to be used as a gold standard. The responses to the question: *What information have you searched for and been unable to find on the university's websites?* showed that 58.26% of users have difficulty finding scholarship announcements, which led us to focus on this type of information. Additionally, we used the responses to the question: *List three questions you would likely ask a virtual assistant about announcements* as the basis for building the gold-standard questions. These questions represent the ideal queries the chatbot should be capable of addressing effectively.

This process resulted in a robust set of 338 questions. With the help of a university staff specialist in student affairs, we crafted a corresponding reference answer for each question. Table 4 exemplifies a subset of elicited questions and their ground truth by category. The complete set is available in our replication package [23].

**Table 4: The subset of questions and their ground truth**

| Category | Question | Ground truth[5] |
|---|---|---|
| Professors (130 questions) | What is the academic background of Professor Dr.[R]? | Bachelor's degree in [Undergraduate course] from the [X University] and [Y University], [Country] [Year]; Master's degree in Software Engineering from [Z University], [City] [Year]; Ph.D. in Computer Science from [B University]. |
| Academic Calendar (18 questions) | What is the start date for the Final Evaluation Period of Semester 2024.1 for distance learning undergraduate programs? | The Final Evaluation Period for Semester 2024.1 for distance learning undergraduate programs begins on July 1, 2024. |
| Enrollment and Procedures (40 questions) | Can incoming students request a partial withdrawal? | Partial withdrawal is not allowed in the semester of entry for the student in the course. |
| Scholarship Announcements (150 questions) | What are the requirements to be a scholarship holder for the XX/YY teaching initiation program? | a) Be regularly enrolled in curricular components of a [XPTO University] undergraduate course (on-site or distance learning) that total at least 12 (twelve) weekly hours during the monitoring period; b) Have 12 (twelve) hours per week available for the monitoring duties; [...] |

## 4.2 Usability Evaluation

To evaluate Freire's usability, we conducted a survey based on the Chatbot Usability Questionnaire (CUQ) [10]. The CUQ is based on the chatbot UX principles provided by the ALMA Chatbot Test tool [22], which assesses the personality, onboarding, navigation, understanding, responses, error handling, and intelligence of a chatbot. The CUQ was designed to be comparable to the System Usability Scale (SUS) [2], but adapted for chatbots and comprises 16 statements. Table 5 shows the survey statements, in which the odd-numbered questions address positive aspects of the chatbot, while even-numbered questions focus on negative aspects. In line with CUQ, we used a 5-point Likert scale [16] to measure participants' levels of agreement with each statement, ranging from *Strongly disagree* to *Strongly agree* and including a neutral value.

**Table 5: Statements used for measuring the chatbot usability**

| ID | Statement |
|---|---|
| 1 | The chatbot's personality was realistic and engaging |
| 2 | The chatbot seemed too robotic |
| 3 | The chatbot was welcoming during initial setup |
| 4 | The chatbot seemed very unfriendly |
| 5 | The chatbot explained its scope and purpose well |
| 6 | The chatbot gave no indication as to its purpose |
| 7 | The chatbot was easy to navigate |
| 8 | It would be easy to get confused when using the chatbot |
| 9 | The chatbot understood me well |
| 10 | The chatbot failed to recognize a lot of my inputs |
| 11 | Chatbot responses were useful, appropriate and informative |
| 12 | Chatbot responses were not relevant |
| 13 | The chatbot coped well with any errors or mistakes |
| 14 | The chatbot seemed unable to handle any errors |
| 15 | The chatbot was very easy to use |
| 16 | The chatbot was very complex |

**Target survey population.** Our target population consists of undergraduate students in different academic periods who meet the following criteria: (1) experience with digital technologies (such as smartphones, tablets, or computers); and (2) some familiarity with chatbots. These criteria ensure that participants have sufficient exposure to digital technologies and experience with chatbot interactions, providing a relevant context for evaluating Freire's usability. Table 6 overviews the participant's characterization.

**Table 6: Participant characterization**

| Characteristic | Category | # Count | Pct. |
|---|---|---|---|
| Age | 18-24 years | 44 | 98% |
| | 25-34 years | 1 | 2% |
| Gender | Man | 33 | 73.3% |
| | Women | 12 | 26.7% |
| Course | Systems Analysis and Development | 21 | 46.7% |
| | Data Science | 12 | 26.7% |
| | Information Security | 12 | 26.7% |
| Academic Period | 1st to 2nd semester | 24 | 53.3% |
| | 3rd to 4th semester | 11 | 24.4% |
| | 5th to 6th semester | 10 | 22.2% |
| Familiarity with Chatbots | Very Familiar | 11 | 24.4% |
| | Familiar | 16 | 35.6% |
| | Moderately Familiar | 14 | 31.1% |
| | Little Familiar | 3 | 6.7% |
| | Not Familiar | 1 | 2.2% |
| Frequency of Chatbot Usage | Daily | 19 | 42.2% |
| | Weekly | 18 | 40% |
| | Monthly | 6 | 13.3% |
| | Rarely | 2 | 4.4% |

**Freire Usage and Survey Execution.** Before conducting the survey, we introduced to each participant the Freire. Specifically, we demonstrated its main purpose and all the functionalities described in Section 3. To illustrate these functionalities, we simulated an interaction with a student seeking scholarship information. Subsequently, we organized a raffle to determine the order in which each participant would use the Freire. Each participant then had an individual session to explore the Freire. Once set up, each participant was free to utilize the Freire in any way they wished, and we encouraged them to explore all the features it offered. We set a five-minute limit for this exploration phase to provide a structured yet flexible timeframe, allowing participants to experience a meaningful interaction without prolonged exposure. After completing their interaction with Freire, we asked the participants to fill out a survey that included the questions presented in Table 5. Additionally, we included two optional open-ended questions: *What positive aspects would you highlight about FREIRE?*, and *What improvement suggestions would you recommend? What did you feel was missing?*. To analyze these questions, we applied the Grounded Theory procedures by performing open and axial codings to improve the reliability of the findings [8] by systematically categorizing responses to identify key themes and patterns.

# 5 Results and Discussion

## 5.1 Contextual Quality of Freire's Responses

Table 7 presents the contextual quality and end-to-end evaluation results across four categories: Professors, Academic Calendar, Enrollment and Procedures, and Scholarship Announcements. The evaluation is split into three components: Retrieval, Generation, and End-to-End Evaluation. Retrieval, which evaluates the efficiency of the retrieval process using Context Precision (Prec.), Context Recall (Rec.), and Context Relevance (Rel.). Generation, which measures the contextual relevance of the generated responses with Faithfulness (Faith.), and Relevance (Rel.). Finally, the End-to-End Evaluation, which measures different dimensions of Freire's responses regarding user experience through Coherence, Clarity, Context Appropriateness, and Instruction Following (Inst. Following). Tables 2 and 3 provide the full descriptions of these metrics.

In Table 7, we can observe that the **Professors** category obtained the best results across most metrics. Particularly, in Retrieval (Context Prec. = 0.90, Context Rec. = 0.86, Context Rel. = 0.90) and End-to-End Evaluation (Coherence = 0.91, Clarity = 0.86, Context Appropriateness = 0.89, Inst. Following = 0.82). These results demonstrate that Freire effectively addresses questions in this category, providing responses that are not only highly relevant to the context but also exhibit strong coherence and clarity.

In the **Academic Calendar** category, we can observe that Freire performed reasonably well, particularly in Generation (Faith. = 0.78, Rel. = 0.82), and Retrieval (Context Prec. = 0.81, Context Rec. = 0.76). However, in the End-to-End Evaluation, Freire obtained an intermediate performance, with Instruction Following scoring 0.75, slightly below the ideal. Despite this, Freire maintained high levels of Coherence and Clarity, both at 0.86, indicating that the responses are clear, consistent, and easy to understand. This suggests that while Freire may face challenges in following instructions precisely, it still provides responses that are logically structured and comprehensible. Similarly, in the **Enrollment and Procedures** category, Freire showed slightly lower scores in Context Recall (0.74) and

**Table 7: Overviews the results of the contextual quality and end-to-end evaluation**

| Category | Retrieval | | | Generation | | End-to-End Evaluation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Context Prec. | Context Recall | Context Rel. | Faith. | Rel. | Coherence | Clarity | Context Appropriateness | Inst. Following |
| Professors | 0.90 | 0.86 | 0.90 | 0.85 | 0.91 | 0.91 | 0.86 | 0.89 | 0.82 |
| Academic Calendar | 0.81 | 0.76 | 0.82 | 0.78 | 0.82 | 0.86 | 0.86 | 0.84 | 0.75 |
| Enrollment and Procedures | 0.76 | 0.74 | 0.77 | 0.75 | 0.78 | 0.88 | 0.87 | 0.81 | 0.75 |
| Scholarship Announcements | 0.77 | 0.72 | 0.79 | 0.73 | 0.78 | 0.90 | 0.86 | 0.78 | 0.71 |

Faithfulness (0.75), suggesting challenges in accurately retrieving and reproducing relevant information. However, it maintained a high Coherence score (0.88), demonstrating that the responses remained logically consistent and understandable. This highlights Freire's ability to produce clear, coherent answers, even when the accuracy of the retrieved information could be improved.

The **Scholarship Announcements** category obtained the lowest scores overall, particularly in Retrieval (Context Recall = 0.72), Generation (Faithfulness = 0.73), and End-to-End Evaluation (Instruction Following = 0.71, Context Appropriateness = 0.78). These results can be attributed to the varied nature of scholarship announcements, as they often contain different types of information, such as eligibility criteria, deadlines, and application processes, that require distinct handling. The diversity in content can challenge the system's ability to retrieve and generate contextually appropriate responses according to each announcement.

Additionally, the high volume of tokens (218.671) in these documents is likely related to this issue, as processing such a large and diverse dataset can make it difficult for the model to maintain coherence and relevance. In the RAG context, this presents a challenge because the model must navigate a broader range of information, which increases the likelihood of retrieving less relevant or conflicting details. As a result, the system's ability to provide accurate, contextually appropriate, and faithful responses is compromised. This underscores the complexity of processing diverse and extensive datasets, particularly in domains with varying content structures.

In summary, Freire outperformed in the Professors category, demonstrating high levels of Retrieval accuracy and End-to-End Evaluation quality, with strong coherence and clarity. In contrast, the Scholarship Announcements category showed more challenges, with lower performance across retrieval, generation, and end-to-end metrics. This can be attributed to the diverse nature of scholarship announcements, which involve varying types of information that complicate the model's ability to provide consistently relevant and accurate responses. Despite these challenges, Freire generally maintained high coherence and clarity, demonstrating its strength in delivering clear, logically structured responses.

**$RQ_1$ Summary:** The Freire Assistant outperformed in the Professors category, but faced challenges when handling more complex categories, such as scholarship announcements. Despite this, all categories maintained consistently high levels of Coherence and Clarity ($\geq 0.86$). This reveals Freire's robustness in ensuring logical consistency and clarity, even in more challenging scenarios.

## 5.2 Students' Perception of Freire's Usability

Figure 3 illustrates the distribution of CUQ (Chatbot Usability Questionnaire) scores across participants. The x-axis represents individual participants, and the y-axis shows their respective CUQ scores on a scale from 0 to 100. Each point reflects the usability score given by a participant, allowing for a clear visualization of the variation in user experiences with the chatbot.

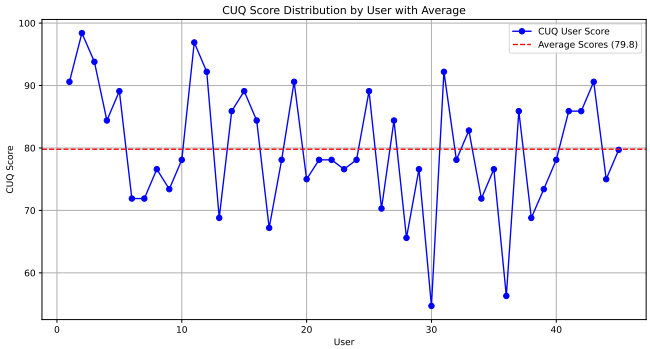


**Figure 3: CUQ Score distribution by participant with average**

**The distribution of the CUQ score.** By looking at Figure 3, we can observe that the distribution of CUQ scores indicates that, on average, users rated Freire's usability positively, with an average score of 79.8. However, there is notable variability in individual scores, indicating mixed user experiences. While many participants scored at or above the average, a subset provided lower scores, suggesting potential usability challenges for certain students, which could suggest areas for targeted improvements. For instance, P30, who reported limited experience with chatbots, might have faced challenges impacting their score, underscoring how familiarity can influence usability perceptions. To further understand the possible issues, we analyzed the average scores of each question as follows.

Figure 4 presents the average scores for the survey questions that assess Freire's usability. Each bar represents the mean score for a specific question, and the error bars indicate the variability (standard deviation) of the responses. The x-axis represents the question numbers, ranging from 1 to 16 according to Table 5. Odd-numbered questions are positive statements about the chatbot, while even-numbered questions are negative. The y-axis indicates the average score on a scale from 0 to 5. Positive questions (odd-numbered) are colored blue, and negative questions (even-numbered) are colored red, providing a visual contrast between the statements.

Positive statements consistently received higher scores, with blue bars generally higher than red bars, indicating a favorable student perception of the chatbot's attributes. Specifically, in terms

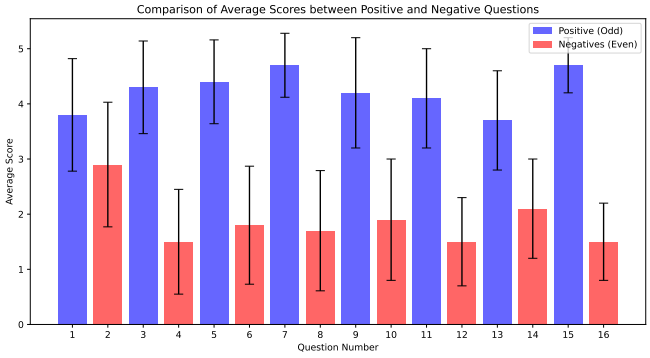


**Figure 4: Average scores for positive and negative questions**

of Freire's **personality**, Statement 1, *The chatbot's personality was realistic and engaging*, achieved a high score (3.1 ± 0.9). This suggests that students found the chatbot's personality authentic and engaging, contributing positively to the user experience. Concerning the **navigation** (Statement 7) and **usability** (Statement 15), the Freire also received high average scores (4.8 ± 0.5), and (4.7 ± 0.5), respectively. This indicates that students found the chatbot user-friendly and easy to navigate. Regarding Freire's **understanding**, Statement 9, *The chatbot understood me well*, also received a high score (4.3 ± 0.9), showing that students felt understood by the chatbot. However, the negative counterpart, Statement 10, had a low score (1.9 ± 1.1), further reinforcing the chatbot's perceived effectiveness in understanding inputs. A similar observation is valid to Freire's **responsiveness** (Statements 11 and 12) with 4.2 ± 0.9, and 1.4 ± 0.8, respectively.

The statements on **error handling** showed mixed feedback. Statement 13, *The chatbot coped well with any errors or mistakes*, received a moderate score (3.7 ± 0.8), while, its counterpart, Statement 14, scored lower (2.2 ± 0.9), indicating that, although the Freire manages some errors effectively, there is room for improvement. Finally, about Freire's clarity of purpose, Statement 5, *The chatbot explained its scope and purpose well*, scored high (4.3 ± 0.8), meaning students felt the chatbot clarified its role. Conversely, Statement 16, *The chatbot was very complex*, had a low score (1.4 ± 0.7), indicating students found the chatbot reasonably simple to use.

In summary, the CUQ scores reveal that the positive statements tend to have higher scores, indicating that participants rated the chatbot more favorably on positive attributes like ease of use, realistic personality, and effective understanding. Conversely, negative attributes, such as the chatbot being too robotic or failing to recognize inputs, received lower scores.

**The positive aspects.** Table 8 shows the eight categories of positive aspects mentioned by students in response to the question: *What positive aspects would you highlight about FREIRE?*.

We can observe that Freire can offer **Support and Reception for Students** who may not typically seek help directly as mentioned by P12 *"It is a very promising idea for students who are not used to asking questions directly to someone."*, and P14 *"It's a chat that cleared up my doubts very well"*. This indicates that Freire Assistant provides a more comfortable, indirect way for students to engage and resolve doubts. The students often mentioned the **Attractive Design and Friendly Interface** of Freire. For instance, the P32 stated: *"The design is very well-built and intuitive."*. Additionally, the visual aspects related to accessibility, such as color schemes and the use of dark/light modes appear as positive aspects as mentioned by P28 *"The colors and the alternation between dark and light mode help a lot of people."*. The students also stated about the **Stability and Security** of responses provided by Freire, e.g., P39 stated: *"I tried to make it get lost and hallucinate, but it didn't happen, so I would say it is quite stable.."* This highlights the Freire's stability in providing consistent, accurate responses.

**Table 8: Positive aspects cited by students**

| Category | Subcategory | # Cited |
|---|---|---|
| Support and Reception for Students | Support and Reception for Students | 3 |
| | Doubt Resolution Capability | 3 |
| Attractive Design and Friendly Interface | Friendly Design and Interface | 11 |
| | Visual Accessibility | 2 |
| | Easy Navigation | 1 |
| Stability and Security | Information Reliability | 1 |
| | Stability of Responses | 1 |
| Informativeness and Relevance | Relevance of Provided Information | 4 |
| Intuitiveness and Ease of Use | Ease of Use | 6 |
| | Intuitiveness and Ease of Use | 3 |
| | Interactivity | 1 |
| Accessible Language | Easy-to-Understand Language | 4 |
| | Objectivity and Simplicity in Responses | 4 |
| | Clarity and Relevance of Responses | 2 |
| | Description and Elaboration of Responses | 1 |
| | Precision and Completeness in Responses | 2 |
| Practicality of Use | Practical Use | 6 |
| Response Speed | Response Speed | 2 |

The students also mentioned positive aspects related to **Informativeness and Relevance** of the Freire responses, as stated by P7: *"It is very effective in providing information and giving instructions"*, indicating that the assistant meets its purpose of assisting students with clear and actionable information. Many students mentioned the **Intuitiveness and Ease of Use**. For instance, P3 and P2 stated: *"I found it simple and easy to use"* and *"I found it very intuitive and easy to interact with."*, respectively. Its intuitive interface and interactive features contribute to a seamless user experience. The **Accessible Language** used by the assistant is clear and easy to understand, as evidenced by P3 *"[...] The language it uses is also easy to understand."*. This allows students to interact with the assistant in a more accessible way, receiving direct and concise answers without confusion. Finally, the **Practicality of Use** was highlighted for its ability to save time in searching for information, as stated by P2: *"a great tool for clarifying doubts about University, [..] very practical, a way to save time and be more direct in what the student needs to do."*.

In summary, the Freire Assistant received positive feedback from students. Its ease of use, effective support, clear communication, and practical functionality have contributed to a positive experience. Additionally, its intuitive design and accessible language made it easy to use and understand, while its practicality helped students save time in finding the information that they needed.

**The suggestions of improvement.** Table 9 shows the suggestions for improvement mentioned by students in response to the question: *What improvement suggestions would you recommend? What did you feel was missing?*.

**Table 9: Suggestions for improvement cited by students**

| Category | Subcategory | # Cited |
|---|---|---|
| Update, Expansion, Accuracy, and Direction of Information | Expansion of Information | 5 |
| | Information Updates | 2 |
| | Precision of Information | 1 |
| | Clarity of Information | 1 |
| | Specific Direction | 1 |
| Feature Request on Conversations | Conversation History | 2 |
| | Topic Grouping | 1 |
| | FAQ Mode | 1 |
| Humanization and Simplicity of Responses | Response Humanization | 3 |
| | Response Length | 1 |
| | Simplicity of Responses | 1 |
| Specificity of Responses | Specificity of Responses | 4 |
| Response Delay | Response Delay | 12 |
| Navigability and Interaction | Opening Links in a New Tab | 1 |
| | Interaction Experience with Voice Recognition | 1 |
| | Button Placement | 1 |
| | Visual Spacing of Responses | 1 |
| Real-Time Visualization and Processing Indicator | Response Transition and Loading | 3 |

While the Freire is generally informative, some students mentioned opportunities for improvement. Some felt that Freire's information could be more detailed and up-to-date. For instance, P23 suggested: *"[It could] have more information about complementary and extension hours."*. Other students pointed out vague information, with P10 stating: *"Update the system's data [...] some information is very vague."*. Additionally, the need for more specific directions was emphasized by P17: *"[...] Implement the ability to specify which UFC campus you want as the basis for answers."*. Regarding the response time and delays, although students generally found the assistant helpful, many complained about slow response times. For instance, the P39 mentioned: *"I found its response time a bit slow."*. This suggests that, while the assistant provides useful responses, the delay may hinder the user experience.

Furthermore, while Freire's language was accessible, some students expressed a desire for more informal, human-like responses, P7 suggested *"More informality; it creates connection and comfort."*. This indicates that a more conversational tone could enhance user satisfaction. Regarding simplicity and conciseness, while the assistant was praised for clarity, some students suggested further simplification. P2 commented: *"Try to make the answers shorter or more concise."* Others expressed a preference for simpler responses overall, as stated by P44: *"A simplicity in the responses."* This shows that while the responses are clear, some students may benefit from even more simplified communication.

Additionally, while the assistant's interface was appreciated for its simplicity, some students suggested improvements in layout and navigability. For instance, P38 suggested: *"Place the user's question bubble on the right side of the screen and the bot's response on the left."*. Similarly, P30 highlighted the need for improved visual organization: *"Responses with exaggerated spacing, making reading difficult."*. Finally, the students also suggested adding more features to improve functionality. For instance, P6 suggested adding a history feature to save past conversations: *"I recommend creating a history to save past conversations [...]."*. Additionally, P28 suggested a better organization of conversations: *"[...] It is necessary to create several 'chat stations' for easy access and grouping of conversations with the chatbot."* Finally, a FAQ module was suggested by P39: *"FAQ module for the user to understand what the chatbot is capable of doing."*

**RQ2 Summary:** Freire Assistant provides clear and useful information while offering an intuitive and accessible user experience. However, opportunities for improvement regarding accuracy and response speed, personalization, and interaction flow are essential to increase its effectiveness and user satisfaction.

## 6 Limitations and Lessons Learned

Despite the promising results, we identified limitations that offer opportunities for improvement in similar future studies. We list some limitations and lessons learned as follows.

**Response Time Delays.** During the qualitative evaluation, response delays (15-20s) were identified as a frequent issue. This issue resulted from a design decision aimed at enhancing query processing by refining the user's input query through the LLM. While this approach resulted in more accurate responses, it also increased the time required to generate them. To address this limitation, future iterations could explore the implementation of streaming techniques, which would improve user experience by delivering partial responses in real time. Additionally, employing strategies such as using emoticons and reactions could help reduce the perception of delays, maintaining user engagement during response generation.

**Outdated or Missing Information.** The qualitative analyses also revealed issues related to outdated or incomplete information provided by the chatbot. This issue primarily arises from the difficulty in obtaining up-to-date public notices and accessible data about the University. To address this limitation, future iterations will require enhanced data collection mechanisms and strategies are needed to ensure the timely updating of relevant content, ensuring the chatbot delivers accurate and current information.

**Hallucination Cases**. In certain situations, the chatbot exhibited hallucinations, especially when the conversation topic shifted within the same context. To mitigate this, future enhancements could include the integration of a semantic router during document uploads. The semantic router works by routing user input to the appropriate response mechanisms based on the contextual meaning of the query. This enables better control over handling queries and ensures the search mechanism retrieves more precise and contextually relevant responses.

## 7 Threats to Validity

We discuss threats to the study validity [33] as follows.

**Construct and Internal Validity.** The selected metrics to conduct the end-to-end and contextual quality evaluation of the chatbot, may not fully capture the users' overall satisfaction or the chatbot's true effectiveness in real-world scenarios. However, we selected the most appropriate and comprehensive metrics available to measure the chatbot's performance [28]. Additionally, internal validity may be affected by potential biases in the survey responses, which could arise from the students' familiarity with the chatbot or their previous experiences with similar technologies. However, we believe that these biases are minimized by using a diverse participant

sample and ensuring that the chatbot's design is intuitive for both novice and experienced users. **Conclusion and External Validity.** We conducted a single case study at one Brazilian university, which may limit the ability to draw definitive conclusions about the Freire Assistant's effectiveness. Additionally, We have partially relied on our data analysis protocol on GT [8]. We aimed to reduce the inherent subjectivity of the coding process. Thus, we analyzed all data in a pair to minimize biases and reach a consensus about the categories and subcategories. Regarding external validity, the specific demographic and institutional context of the case study may not be representative of all universities, potentially limiting the generalizability of the findings to other educational institutions, regions, or even different chatbot implementations.

## 8 Conclusion and Future Work

This paper presents the Freire Assistant, a smart chatbot powered by LLM and RAG pipelines to improve access to academic opportunities and institutional matters. We evaluated it based on response quality and relevance, and overall usability. The Freire Assistant demonstrates significant potential in helping students find academic information by providing a user-friendly, intuitive, and reliable platform. It outperformed with strong contextual relevance, coherence, and clarity in providing responses related to the Professors category, and performed well in more complex areas like the Academic Calendar, and Enrollment, though challenges were noted in information retrieval and instruction following.

For future work, we envision several directions to improve and expand Freire Assistant: (i) *Incorporating Additional Information Categories* such as events, etc. This would allow Freire to meet a broader spectrum of student needs; (ii) *Adapting to Other Academic Unit* ensuring adaptability to different academic contexts, regulations, and administrative structures. This would involve collecting and processing data from other institutions while maintaining Freire's focus on user experience and contextual relevance; and (iii) *Implementing Semantic Routing Strategies* to enhance information retrieval. These strategies aim to direct queries to specific document categories or models based on the semantics of the query, improving response accuracy and efficiency.

**Acknowledgments.** This study was partially financed by UFC-PIBI and FUNCAP (BP5-00197-00042.01.00/22).

## References

[1] Suha Khalil Assayed, Khaled Shaalan, and Manar Alkhatib. 2022. A chatbot intent classifier for supporting high school students. *EAI Endorsed Transactions on Scalable Information Systems* 10, 3 (2022).
[2] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.
[4] Victor R Basili-Gianluigi Caldiera and H Dieter Rombach. 1994. Goal question metric paradigm. *Encyclopedia of software engineering* 1, 528-532 (1994), 6.
[5] Yogi Wisesa Chandra and Suyanto Suyanto. 2019. Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model. *Procedia Computer Science* 157 (2019), 367–374.
[6] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access* 8 (2020), 75264–75278.
[7] Yu-Hung Chien and Chun-Kai Yao. 2020. Development of an ai userbot for engineering design education using an intent and flow combined framework. *Applied Sciences* 10, 22 (2020), 7970.
[8] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage publications.
[9] Mozilla Corporation. 2024. Web Speech API. https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API. (Accessed on 17/10/2024).
[10] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?. In *31st European Conference on Cognitive Ergonomics*. 207–214.
[11] Confident AI Inc. 2024. Deepeval documentation. https://docs.confident-ai.com/. (Accessed on 17/10/2024).
[12] Jina. 2024. Jina framework. https://jina.ai/reader/. (Accessed on 17/10/2024).
[13] langchain. 2024. Langchain framework. https://www.langchain.com/. (Accessed on 17/10/2024).
[14] Langchain. 2024. RecursiveCharacterTextSplitter langchain documentation. https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/
[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[16] R. Likert. 1932. *A Technique for the Measurement of Attitudes.* Number Nº 136-165 in A Technique for the Measurement of Attitudes. Archives of Psychology.
[17] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[18] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *55th ACM Technical Symposium on Computer Science Education V. 1*. 750–756.
[19] Bei Luo, Raymond YK Lau, Chunping Li, and Yain-Whar Si. 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 1 (2022), e1434.
[20] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.
[21] C Martinez-Araneda, M Gutiérrez, D Maldonado, P Gómez, A Segura, and C Vidal-Castro. 2024. Designing a Chatbot to Support Problem-Solving in a Programming Course. In *INTED2024 Proceedings*. IATED, 966–975.
[22] Ábalos N. Martín J., Muñoz-Romero C. 2024. chatbottest - Improve your chatbot's design. https://chatbottest.com/. (Accessed on 17/10/2024).
[23] Arthur Willame Mesquita, Carlos Freire, Antonio Gomes, Bruna Amazonas, and Anderson Uchôa. 2025. *Replication package for the paper: "Hello, Freire! How Can You Help Me? A Smart Chatbot to Enhance Access to Academic Opportunities and Institutional Matters".* https://doi.org/10.5281/zenodo.15014430
[24] Subash Neupane, Elias Hossain, Jason Keith, Himanshu Tripathi, Farbod Ghiasi, Noorbakhsh Amiri Golilarz, Amin Amirlatifi, Sudip Mittal, and Shahram Rahimi. 2024. From Questions to Insightful Answers: Building an Informed Chatbot for University Resources. *arXiv preprint arXiv:2405.08120* (2024).
[25] Pedro Filipe Oliveira and Paulo Matos. 2023. Introducing a chatbot to the web portal of a higher education institution to enhance student interaction. *Engineering Proceedings* 56, 1 (2023), 128.
[26] OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4
[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics*. 311–318.
[28] Anupam Purwar et al. 2024. Evaluating the Efficacy of Open-Source LLMs in Enterprise-Specific RAG Systems: A Comparative Study of Performance and Scalability. *arXiv preprint arXiv:2406.11424* (2024).
[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.
[30] J Jinu Sophia and T Prem Jacob. 2021. Edubot-a chatbot for education in covid-19 pandemic and vqabot comparison. In *2nd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 1707–1714.
[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[32] Yujia Wang, Shuqi Liu, and Linqi Song. 2022. Designing an educational chatbot with joint intent classification and slot filling. In *International Conference on Teaching, Assessment and Learning for Engineering (TALE)*. IEEE, 381–388.
[33] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.
[34] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469* (2023).