ISSN: 2594 - 2905

# DEVELOPMENT OF CHATBOT FOR THE TECHNOLOGICAL INSTITUTE OF ZACATEPEC

DEVELOPMENT OF CHATBOT FOR THE TECHNOLOGICAL INSTITUTE OF ZACATEPEC

Pérez Calderón Carlos[1] , Vázquez Carrillo Luis Ángel[2] , Muñiz-Rascado Luis José[3] ,
Velázquez-Santana Eugenio César[4] , Castro-Sánchez Noé Alejandro[5]

https://doi.org/10.61117/ipsumtec.v8i2.390

[1] National Technological Institute of Mexico/Technological Institute of Zacatepec,L19091404@zacatepec.tecnm.mx , https://orcid.org/0009-0003-0590-4885.
[2] National Technological Institute of Mexico/Zacatepec Technological Institute,L20091199@zacatepec.tecnm.mx , https://orcid.org/0009-0007-9101-5513.
[3] National Technological Institute of Mexico/Zacatepec Technological Institute,luis.mr@zacatepec.tecnm.mx , https://orcid.org/0000-0003-2941-2482.
[4] National Technological Institute of Mexico/Zacatepec Technological Institute,eugenio.vs@zacatepec.tecnm.mx , https://orcid.org/0000-0002-7683-7224.
[5] National Technological Institute of Mexico/CENIDET,noe.cs@cenidet.tecnm.mx , https://orcid.org/0000-0002-8083-3891.

***Abstract*** – This article presents the development of a chatbot for the ITZ Zacatepec Institute of Technology, trained with relevant data from the institution to provide agile and natural responses to user queries on the university's website. The proposed chatbot uses *large language model (LLM*) technology, together with advanced information retrieval algorithms. The system seeks to reduce the administrative burden on staff and offer users an interactive and efficient experience. The end result is a computer system that receives user queries; the BM25 retrieval algorithm searches a corpus of data related to the institution, and finally the LLM model processes it, generating a response to the request.

***Keywords:*** BM25, Chatbot, AI, LLM.

**Abstract** -- The present article presents the development of a chatbot designed for TecNM Instituto Tecnológico de Zacatepec, trained with relevant institutional data to provide agile and natural user assistance on the university's website. The proposed chatbot utilizes Large Language Model (LLM) technology along with advanced information retrieval algorithms. The system aims to reduce the administrative burden on staff and offer users an interactive and efficient experience. The result is a computational system that receives user queries; the BM25 retrieval algorithm searches through a corpus of institution-related data, and finally, the LLM model (GEMMA) processes the context, generating an appropriate response to the request.

***Keywords:*** Chatbot, AI, LLM, BM25.

## INTRODUCTION

Traditional customer service methods, such as telephone lines and emails, are insufficient to handle a large volume of inquiries, resulting in long wait times and a considerable administrative burden. However, artificial intelligence (AI) has now taken on a fundamental role in optimizing these processes. According to Caldarin et al. [1], AI encompasses various areas such as natural language processing (NLP), speech synthesis, and speech comprehension, among others. Furthermore, its ability to systematize and analyze large volumes of information makes it a key tool for task automation.

The rise of AI has driven automation in multiple sectors. A chatbot is a computer program that simulates human conversation with an end user, according to IBM [15]. Chatbots potentially represent a new paradigm in the way people will interact with data and services in the future [12], facilitating user support and optimizing information management.

In this context, the development of a chatbot was proposed for the TecNM Instituto Tecnológico de Zacatepec, with the aim of providing a virtual assistant that answers questions and provides institutional information. This will reduce the administrative workload and offer quick and efficient responses to users.

## DEVELOPMENT

The development of this project began with qualitative research, in which several LLM models were investigated as technology to power the chatbot. These models underwent an evaluation process to determine which was the most optimal for the project. Next, a search algorithm was implemented to locate relevant documents within a corpus of

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

documents, to be used by the model to generate responses to questions using the retrieved information as context.

**BM25 retrieval algorithm.**

This algorithm was proposed by two researchers [3], a British programmer with a degree in mathematics from the University of Cambridge, and Spärck Jones, a British scientist specializing in computational linguistics. They developed a mathematical model based on probabilistic principles for document localization. Their approach focuses on determining the relevance of certain words within a corpus of documents, dividing the process into two fundamental calculations: TF and IDF.

**TF.**

Measures the relevance of a term within a document, considering the number of times appears in that document, although it may be biased in long documents simply because they have more words. That is why, unlike the approaches simple as TF-IDF, BM25 uses a "normalized" version of TF to handle better situations in which a term appears very frequently to avoid this bias, this calculation is obtained from equation Ec.(1).

$$TF = \left[ \frac{f(q_i, D) \cdot (k_i + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \right] \quad \text{Eq. (1)}$$

Where:

- $f(q_i, D)$: Indicates how many times the term appears $q_i$ appears in the document $D$.
- $k_i$: This is an adjustment parameter that controls the sensitivity of term frequency.
- $k_i + 1$ increases the sensitivity to give more weight to the term frequency.
- $b$: Adjustment parameter that controls the length of the document in the formula $b$ takes values between 0 and 1.
- $|D|$: Total number of words in the document.
- $avgdl$: The average length of the documents in the corpus.

**IDF**

*Inverse Document Frequency* (IDF) is a measure used to evaluate the importance of a term within a corpus of documents. IDF assigns a higher weight to terms that are rare in the corpus and a lower weight to terms that are common. The idea is that common terms (such as "the," "a," "of," among others) do not provide much information to distinguish between documents, while rare terms (such as "algorithm," "programming," "neural") are often more relevant for identifying the content of a document.

The classic formula is represented by the following equation Ec.(2).

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad \text{Ec.(2)}$$

Where:

- $t$ is the term or word.
- $N$ is the total number of documents in the corpus.
- $df(t)$ is the number of documents in which the term appears.

**Final combination.**

By combining these two concepts, TF and IDF, we obtain the BM25 algorithm as a result, with 25 coming from the version used for this project [3] Ec.(3).

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \left[ \frac{f(q_i, D) \cdot (k_i + 1)}{f(q_i, D) + k \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \right] \quad \text{Eq. (3)}$$

Where:

- $score(D, Q)$: This is the total relevance score of the document $D$ with respect to the query $Q$.
- $\sum_{i=1}^{n} IDF(q_i)$: Here, the contribution of each $q_{term\,(1)}$ from the query $Q$ to the final score.
- $n$ This is the total number of terms in the query.
- $IDF(q_i)$ is the "Inverse Document Frequency."
- $f(q_i, D)$: This is the Term Frequency (TF).
- $k_i$: This is a tuning parameter that controls the sensitivity of the frequency of the term $f(q_i, D)$.
- $b$: This is another adjustment parameter that controls the length of the document in the formula.
- $|D|$: This is the length of the document $D$, the total number of words in the document.
- $avgdl$: This is the average length of the documents in the corpus.

**Text readability indices:**

These were calculations applied to the text to assess the complexity of both the responses generated by the LLM and the questions asked by the user.

**Fernandez-Huerta.**

Readability refers to the linguistic legibility of the text, i.e., whether it is easy or difficult to understand. It does not cover typographical aspects that greatly influence readability [2], see Eq. (4).

$$L = 206.84 - 0.60P - 1.02F \quad \text{Eq. (4)}$$

Where:

- $L$ = readability.
- $P$ = average number of syllables per word.
- $F$ = average number of words per sentence.

The result is interpreted in Table 1.

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

ISSN: 2594 - 2905

**Table 1**. *Fernández Huerta readability scale.*

| (L) | Grade | School grade |
|---|---|---|
| 90-100 | Very easy | 4th grade |
| 80-90 | Easy | 5th grade |
| 70-80 | Somewhat easy | 6th grade |
| 60-70 | Normal | 7th or 8th grade |
| 50-60 | Somewhat difficult | Pre-university |
| 30-50 | Difficult | Selective courses |
| 0-30 | Very difficult | University |

**Source:** *Simple readability measures (Fernández Huerta, 1959).*

**INFLESZ**

The INFLESZ readability scale measures the ease of reading a text. It was developed by Inés María Barrio Cantalejo. It is adapted to the current average Spanish reader
[6] see Ec.(5)**.**

$$I = 206.835 - 62.3\left(\frac{S}{P}\right) - \left(\frac{P}{F}\right) \qquad \text{Eq.(5)}$$

Where:

- $I$ = INFLESZ scale.
- $S$ = total number of syllables.
- $P$ = number of words.
- $F$ = number of sentences.

The result is interpreted in Table 2.

**Table 2**. *INFLESZ readability scale.*

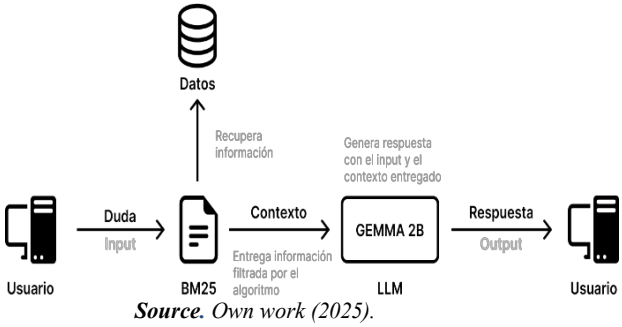
| (L) | Level |
|---|---|
| 80-100 | Very easy |
| 65 | Easy |
| 55 | Somewhat easy |
| 40-55 | Normal (for an adult) |
| 0-40 | Difficult |

**Source:** *Predictive Systems for Written Message Readability (Francisco Szigriszt Pazos, 1992).*

**Application outline.**

The diagram shown in Figure 1 illustrates the architecture of the chatbot system, revealing the synergy between information retrieval algorithms and the use of the GEMMA 2B text processing model.

**Figure 1**. *System architecture.*




***Source.*** *Own work (2025).*

**Evaluation of the models.**

The complexity of the prompts and responses was evaluated using the Fernández-Huerta and INFLESZ indices [2], [6], timing the generation and delivery of responses by the model. **Code 1** shows the Python code used to evaluate the text to determine the results and ascertain the complexity of the *prompts*.

**Code 1.** *Index calculations in Python.*

```
def fernandez_huerta(text): words =
    count_words(text)
    sentences = count_sentences(text)
    syllables = sum(count_syllables(word)
for word in re.findall(r'\b\w+\b'text))

    return 206.84 - 60 * (syllables /
words) - 1.02 * (words/sentences)
```

```
def inflesz(text):
    words = count_words(text)

    sentences = count_sentences(text)
    syllables = sum(count_syllables(word)
for word in re.findall(r'\b\w+\b'text))

    return 206,835 - 62.3 * (syllables /
words) - (words/sentences)
```

To determine the optimal LLM for the project, it was necessary to research the different models available in the sector, with the aim of identifying which one best suits the specific needs in terms of the number of parameters, the consistency of the responses, and the execution time, see Table 3.

**Table 3**. *General information on LLM models.*

| MODEL | PARAMETERS | DEVELOPER |
|---|---|---|
| LLAMA2 | 7B | Goal |
| LLAMA 3 | 70B | Target |
| MISTRAL | 7B | Mistral AI |
| GEMMA | 2B | Google |
| PHI3 | 3B | Microsoft |
| PHI3 | 14 B | Microsoft |
| LLAMA2 | 7B | Meta |

**Source:** *Own elaboration (2025).*

The minimum requirements to run the system are described in Table 4.

**Table 4**. *Initial LLM model requirements.*

| CPU | RAM | VRAM |
|---|---|---|
| Intel Core i5 / AMD Ryzen 5 | 16 GB | 6–8 GB |
| Intel Core i9 / AMD Ryzen 9 | 32 GB | 16 GB |
| Intel Core i7 / AMD Ryzen 7 | 16 GB | 6–8 GB |
| Intel Core i5 / AMD Ryzen 5 | 16 GB | 8 GB |

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

| | | |
|---|---|---|
| Intel Core i5 / AMD Ryzen 5 | 4 GB | 2 GB |
| Intel Core i5 / AMD Ryzen 5 | 8 GB | 4 GB |
| Intel Core i5 / AMD Ryzen 5 | 16 GB | 6–8 GB |

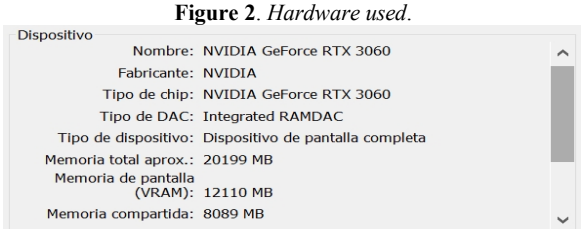
**Source:** *Own elaboration (2025).*

Of these, the following LLM models were selected for desktop testing:

- Gemma 2 trillion parameters.
- Mistral 7 trillion parameters.
- Phi3 3 trillion parameters.
- Llamma2 7 trillion parameters.

The equipment used to evaluate the models had the *hardware* configuration specified in Figure 2. This configuration was considered adequate for the computational demands of the tests.

**Figure 2**. *Hardware used.*




**Source:** *Own work (2025).*

A relationship was established between the complexity of the *prompts* (measured using the Fernández-Huerta and INFLESZ indices). To do this, three levels of complexity were considered: "easy," "normal," and "difficult," according to the scale of the aforementioned indices.

**LLM model configuration.**
In order to ensure the comparability of the results, a uniform configuration was maintained for all LLMs, see **Code 2**:

**Code 2.** *LLM configuration in Python.*

```
outputs = pipeline( prompt,
        max_new_token = 350,
        do_sample = True,
        temperature = 0.7,
        top_k = 50,
        top_p = 0.95.)
```

***Prompts* used in the evaluation.**
For the *prompts* inserted into the LLM models for this test, their complexity indices were based on three levels and four types of questions for each corresponding level:

**Easy:** Easily understandable text that does not require technical studies to be answered. In Table 5.

*Table* **5**. *Easy difficulty prompts.*

| Prompt | INFLES Z scale | Fernández Huerta Scale |
|---|---|---|
| What is the capital of France? | 86.02 | 90.26 |
| Who wrote "One Hundred Years of Solitude"? | 70.45 | 75.26 |
| How many days are there in a leap year? | 70.45 | 75.26 |
| What is photosynthesis? | 68.85 | 73.78 |

**Source:** *Own elaboration (2025).*

**Normal:** text that is understandable and requires prior study to answer; a high school or college student could answer. In Table 6.

**Table 6**. *Prompts with normal difficulty.*

| Prompt | INFLES Z Scale | Fernandez Huerta Scale |
|---|---|---|
| Explain how the theory of relativity works in simple terms. | 56.57 | 61.75 |
| What are the main arguments for and against climate change? | 74.74 | 79.19 |
| Describe the impact of the Renaissance on European art. | 59.74 | 64.87 |
| How do you solve a system of linear equations using matrices? | 45.09 | 50.72 |

**Source:** *Own work (2025).*

**Difficult**: text that is difficult to understand and requires specialized knowledge to answer; a graduate student could answer it. In Table 7.

**Table 7**. *Prompts with difficult difficulty.*

| Prompt | INFLES Z scale | Fernandez Huerta scale |
|---|---|---|
| Analyze the influence of existentialist philosophy on literature . | 0 | 0 |
| Discuss the philosophical and ethical implications of technology from edition | 23.4 | 29.27 |

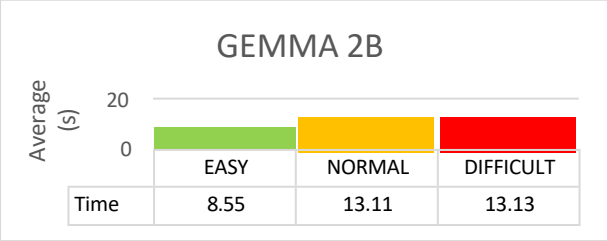
Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

| | | |
|---|---|---|
| Genetics in human evolution. | | |
| Analyze how quantum field theory can be reconciled with the theory of general relativity. | 27.3 | 33.02 |
| Develops a theoretical framework for the integration of artificial intelligence in ethical decision-making and autonomous systems systems. | 24.86 | 30.44 |

**Source:** *Own elaboration (2025).*

**Measurement of execution times.**
Execution times, measured in seconds, were obtained for each model evaluated, considering the complexity of the *prompts* according to the Fernández-Huerta and INFLESZ indices. The results for the GEMMA 2B model are presented in Figure 3.

*Figure* **3**. *GEMMA 2B execution time.*




**Source:** *Own elaboration (2025).*

Table 8 shows the average results of the indices for each level of *prompts* in GEMMA 2B.

*Table* **8**. *Average GEMMA 2B indices.*

| INDEXES | EASY | NORMAL | DIFFICULT |
|---|---|---|---|
| Average Fernández-Huerta index | 78.64 | 64.1325 | 33.145 |
| Average INFLESZ index | 73.9425 | 59.035 | 27.1671 |

**Source:** *Own elaboration (2025).*

The results obtained for the MISTRAL 7B model are presented in Figure 4.

**Figure 4.** *MISTRAL 7B execution time.*

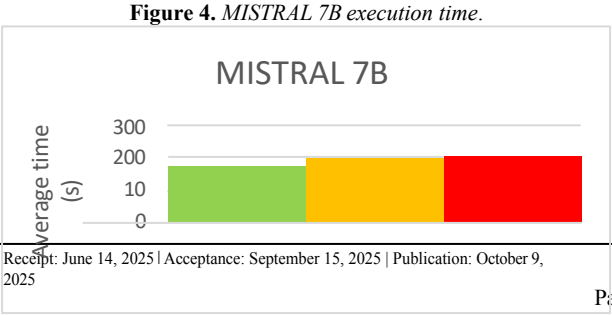



Table 9 shows the average result of the indices for each *prompt* level in MISTRAL.

*Table* **9**. *Average indices in MISTRAL 7B.*

| INDEXES | EASY | NORMAL | DIFFICULT |
|---|---|---|---|
| Average Fernández-Huerta index | 73.4975 | 56.42 | 23.1825 |
| Average INFLESZ index | 68.8925 | 51.2875 | 18.89 |

| | EASY | NORMAL | DIFFICULT |
|---|---|---|---|
| Time | 153.58 | 166.69 | 226.48 |

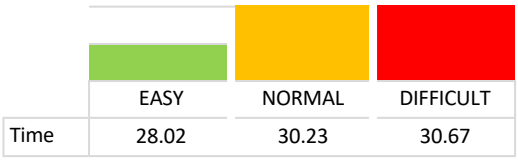
**Source:** *Own elaboration (2025).*

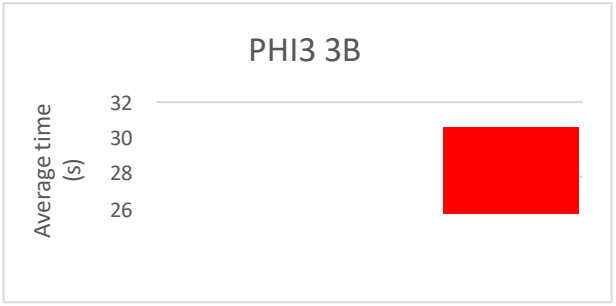
Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

**Source:** *Own elaboration (2025).*

The results obtained for the PHI3 3B model are shown in Figure 5**.**

*Figure* **5**. *PHI3 3B execution time.*




**Source:** *Own elaboration (2025).*

Table 10 shows the average results of the indices for each *prompt* level in PHI3.

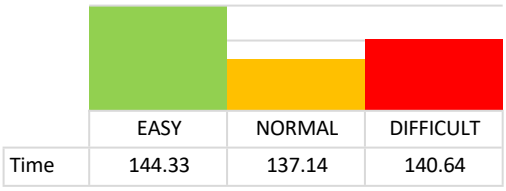
*Table* **10**. *Average PHI3 3B indices.*

| INDEXES | EASY | NORMAL | DIFFICULT |
|---|---|---|---|
| Average Fernández-Huerta index | 73.4975 | 56.42 | 23.1825 |
| Average INFLESZ index | 68.8925 | 51.2875 | 18.89 |

**Source:** *Own elaboration (2025).*

The results obtained for the LLAMA2 7B model are shown in Figure 6**.**

*Figure* **6**. *LLAMA2 7B execution time.*





**Source:** *Own elaboration (2025).*

Table 11 shows the average results of the indices for each *prompt* level in LLAMA2.

*Table* **11**. *Average LLAMA2 7B indices.*

| INDICES | EASY | NORMAL | DIFFICULT |
|---|---|---|---|

IPSUMTEC
ISSN: 2594 - 2905

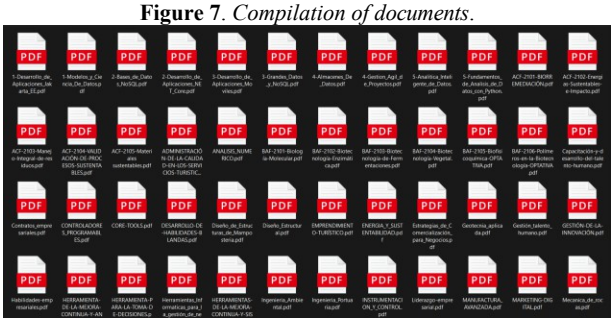
| | | | |
|---|---|---|---|
| Average Fernández-Huerta Index | 100 | 68.95 | 44.7967 |
| Average INFLESZ Index | 98.875 | 65.2825 | 39.3633 |

**Source:** *Own elaboration (2025).*

**Normalization of the document corpus.**

Once the language model was defined, the training phase began. To do this, a corpus of documents from the Zacatepec Institute of Technology was compiled, as shown in Figure 7. These documents will serve as the basis for the model to learn how to generate coherent and relevant responses based on real information.

**Figure 7**. *Compilation of documents.*



**Source:** *Own work (2025).*

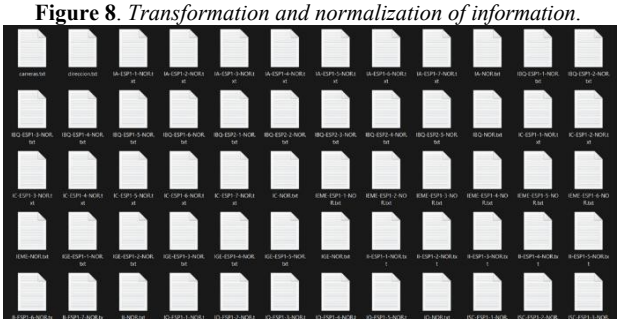
In order to optimize the use of the GEMMA 2B model, a document cleaning process was carried out. This stage consisted of removing special characters, punctuation, and *stopwords* in order to provide the model with a cleaner and more relevant context. **Code 3**.

**Code 3.** *Text cleaning in Python.*

```
def normalizeText(text): #
   Replace characters
    text = text.replace('/', '
    ').replace('-', ' ').replace('_', '')

   # Remove URLs text
   =
   re.sub(r'http\S+|www\S+|https\S+', '',
   text, flags=re.MULTILINE)

   # Remove punctuation text
   = text.
   translate(str.maketrans('', '',
   string.punctuation))
   text = text.lower()

   # Convert to lowercase
   text = unidecode(text)

  # Remove accents
  words = nltk.word_tokenize(text,
  language='Spanish')

  # Tokenize the text
  stop_words =
  set(stopwords.words('Spanish'))
  #Remove stop words
  words = [word for word in words if word
  not in stop_words]
return ' '.join(words)
```

Once normalized and converted to TXT format as shown in Figure 8, the documents were ready to be used as input in the GEMMA 2B model. This flat text format is ideal for training language models.

**Figure 8**. *Transformation and normalization of information.*



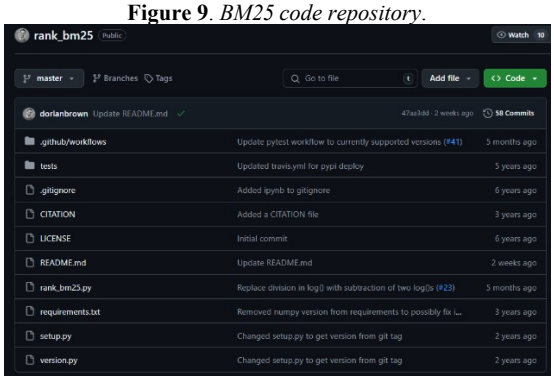
**Source:** *Own creation (2025).*

**Integration of the BM25 search algorithm.**

A code repository available on GitHub was used, hosted at: https://github.com/dorianbrown/rank_bm25.git

as the basis for implementing the algorithm, which is shown in Figure 9. The code provided in the repository was carefully integrated to ensure its compatibility with the system under development.

**Figure 9**. *BM25 code repository.*




**Source:** *GitHub rank_bm25 (Dorian Brown, 2024).*

**BM25 execution integrated with GEMMA 2B.**

To evaluate the functionality of the code, practical tests similar to LLM model evaluations were performed, but focused exclusively on GEMMA 2B with BM25 using ITZ questions.

**Easy:** Its Fernández-Huerta index and INFLESZ index are between 100-70 Table 12.

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

**Table 12**. *Easy prompts with BM25.*

| Prompt | INFLES Z scale | Scale Fernández Huerta |
|---|---|---|
| What is studied in biochemistry? | 100 | 100 |
| What is a database? | 100 | 100 |
| What is done in structural design? | 84.14 | 88.27 |
| What is a slender system? | 93.74 | 89.7 |

**Source:** *Own elaboration (2025).*

**Normal:** Your Fernández-Huerta and INFLESZ indices are between 70-40. Table 13.

**Table 13**. *Normal prompts with BM25.*

| Prompt | INFLES Z scale | Fernández Huerta scale |
|---|---|---|
| What is tourism entrepreneurship? | 52.32 | 57.74 |
| Give me the address of the technology institute. | 52.32 | 57.74 |
| What is electromechanics? | 47.09 | 52.76 |
| Give me the educational program for the bachelor's degree in tourism | 40.3 | 41.18 |

**Source:** *Own elaboration (2025).*

**Difficult:** Su index from Fernández-Huerta as from
INFLESZ is between 40-0 Table 14.

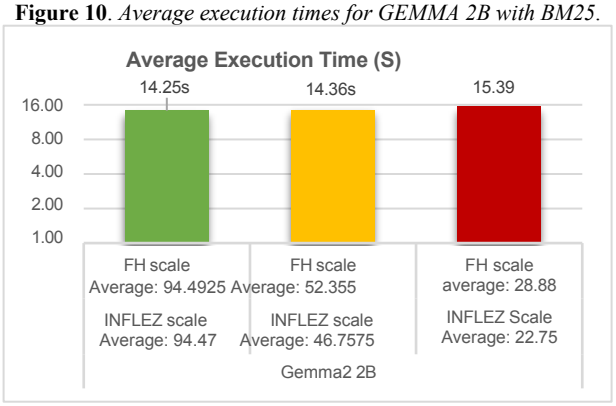
**Table 14**. *Difficult prompts with BM25.*

| Prompt | INFLES Z scale | Fernandez scale Huerta |
|---|---|---|
| What is computer systems engineering? ? | 39.63 | 45.41 |
| Give me the educational program for engineering in administration, specializing in business administration. | 6.94 | 13.58 |
| Give me the educational program for engineering in electromechanics. | 4.15 | 11.18 |
| What is web programming used for and what are its applications in the field of computer engineering? computer systems | 32.28 | 40.35 |

**Source:** *Own elaboration (2025).*

Execution times, measured in seconds, were recorded for GEMMA 2B during the evaluation based on reading the relevant document selected using BM25. In addition to considering the complexity of the prompts using the Fernández-Huerta and INFLESZ indices. The results obtained were averaged, yielding the following values Figure 10.

**Figure 10**. *Average execution times for GEMMA 2B with BM25.*




**Source:** *Own elaboration (2025).*

**API integration**

To enable seamless interaction with the GEMMA 2B model, a RESTful API was designed using the Flask framework. This API exposes an *endpoint* that accepts POST requests with a "*prompt*" field. Upon receiving a request, the model processes the *prompt* and generates a response in JSON format, which is sent back to the client. This can be seen in **Code 4**.

**Code 4**. *Endpoint for requests to the model in Python.*

```
@app.route('/query', methods=['POST']) def
query():
        data = request.json
        query = data.get('query')

        if not query:
        return jsonify({'error': 'No query
provided'}), 400
```

After this process, the query is normalized and tokenized to then calculate the relevance scores with the BM25 algorithm. This process can be seen in **Code 5**.

**Code 5**. *Retrieval of relevant documents with BM25 in Python.*

```
normalized_query = normalize_text(query)
tokenized_query = normalized_query.split()
scores = bm25.get_scores(tokenized_query)

documents_with_scores =
list(zip(filenames, scores))
```

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

```
documents_with_scores.sort(key=lambda x:
x[1], reverse=True)
n_top_documents = 1 top_documents
=
documents_with_scores[:n_top_documents]
```

Once the relevant document has been selected, the GEMMA 2B model is used to generate a summary, as shown in *Code 6*.

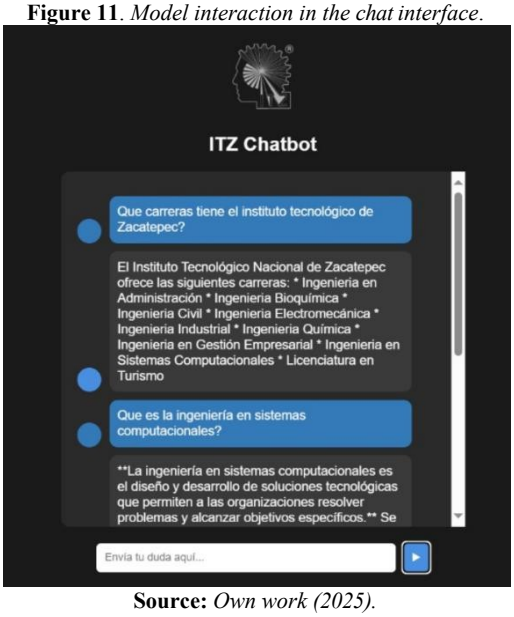
*Code 6.* Generating the summary with GEMMA 2B in Python.

```
messages = [
   {"user",
    "content": f"Summary of the document
    about {query}: {doc_content}"}]

prompt = tokenizer.apply_chat_template(
       messages,
       tokenize = False,
       add_generation_prompt = True)

outputs = gemma_pipeline(
       prompt, max_new_tokens = 256,
       add_special_tokens = True, do_sample
       = True,
       temperature = 0.7,
       top_k = 50, top_p = 0.95)
summary =
  outputs[0]["generated_text"][len(prompt)
:]
```

**Results.**

To evaluate and demonstrate the capabilities of the GEMMA 2B model, a simple user interface was developed using *HTML*, *CSS*, and *JavaScript* web technologies. This interface, which simulates a real-time conversation, allows users to send text messages to the model and receive dynamically generated responses. This interface can be seen in Figure 11.

**Figure 11**. *Model interaction in the chat interface.*




**Source:** *Own work (2025).*

## DISCUSSION AND ANALYSIS OF RESULTS

The combination of large-scale language models (LLMs), such as GEMMA 2B, with information retrieval algorithms such as BM25 proved to be highly effective in resolving questions whose structure does not follow a linear sequential order. This approach made it possible to identify relevant text fragments scattered throughout the base documents, which improved the chatbot's ability to provide coherent and accurate responses. This coincides with the findings reported by Hugo (2009) [14], who delves into the probabilistic bases that make BM25 a robust relevance framework for document retrieval, which theoretically supports the results observed in this research.

Parameters play a crucial role in the performance of a language model (LLM). As Belcic and Stryker [8] explain, in the development of this project, it was found that while a larger number of parameters can improve performance, the associated computational cost increases significantly. Therefore, it is essential to find a balance between model size and available resources, optimizing the number of parameters without sacrificing the quality of the results.

A relevant finding of this research was the significant reduction in the administrative burden resulting from the implementation of the chatbot in automating responses to frequently asked questions, intelligently directing users to relevant sections of the system, and the model's ability to maintain contextualized conversations, which reduced the need for human intervention in repetitive tasks. This observation coincides with the findings of Misischia et al. (2022) [10], in whose study the authors point out that chatbots perform key functions such as problem solving and personalization of service, which directly contributes to improving service quality and optimizing human resources.

A third aspect identified during the development of the chatbot was the potential of generative artificial intelligence to expand the system's capabilities, allowing it to respond to more complex queries and dynamically adapt to the user's conversational style. This technological evolution has proven particularly useful in improving the self-service experience for users by reducing the need to manually program predefined responses and instead generating contextualized responses based on an institutional knowledge base. According to IBM (2023) [15], next-generation chatbots not only replicate human responses, but can generate new content, interpret natural language more accurately, and respond with empathy, significantly improving the quality of interaction. This advancement has allowed to automate processes

Receipt: June 14, 2025 | Acceptance: September 15, 2025 | Publication: October 9, 2025

that were previously handled manually, optimizing staff time and improving overall service efficiency. In addition, the chatbot's ability to understand and address a wider variety of questions has contributed to expanding its usefulness as an institutional service tool.

## CONCLUSIONS

The chatbot at the Zacatepec Institute of Technology represents a breakthrough in the application of artificial intelligence in the academic field. Using the GEMMA 2B model and the BM25 algorithm as a data retrieval method, it improves user interaction and reduces the administrative burden. Evaluation using linguistic complexity indices has optimized its performance, validating its effectiveness in process automation. This project lays the foundation for future AI applications in education.

In relation to these findings, Labadze et al. (2023) [5] conducted a systematic review of 67 studies on educational chatbots, concluding that they offer benefits such as personalized student assistance, administrative support, and skill development, which was demonstrated in this project with its potential to reduce workload. On the other hand, Misischia et al. (2022) [10], within the framework of the 13th International Conference on Environmental Systems, Networks, and Technologies (ANT 2022), analyzed the impact of chatbots on customer service quality, concluding that they significantly improve efficiency, immediacy, and personalization in customer service on digital platforms.

## ACKNOWLEDGMENTS

We express our sincere gratitude to Professor Luis José Muñiz Rascado for his guidance and mentorship throughout the development of this project. His knowledge of generative artificial intelligence and large-scale language models was fundamental to the understanding and application of these concepts in our work. His guidance and advice allowed us to face challenges with a clearer and more informed perspective.

We would also like to extend our gratitude to Professor Eugenio César Velázquez Santana for his constant support and collaboration in the project development process. His help was key not only in structuring and improving the work, but also in the necessary steps for its publication. His willingness and commitment were essential in achieving the objectives set.

## BIBLIOGRAPHY

[1] Caldarin J, Jaf A, McGarry B. The role of chatbots in Industry 4.0 [Internet]. [accessed July 20, 2025]. Available at: https://www.researchgate.net/publication/373718439_The_role_of_Chatbots_in_Industry_40

[2] Fernández Huerta A. Readability [Internet]. Madrid: Legible; [accessed July 20, 2025]. Available at: https://legible.es/blog/lecturabilidad-fernandez-huerta/

[3] Steen H, Urban E. BM25 relevance score configuration [Internet]. Microsoft; 2024 [accessed 20 July July July 2025]. Available at: https://learn.microsoft.com/es-es/azure/search/index-ranking-similarity

[4] Reznik L. General principles and purposes of computational intelligence. Systems Science and Cybernetics [Internet]. 2009 [cited July 20, 2025]. Available from: https://www.eolss.net/sample-chapters/c02/E6-46-04-01.pdf

[5] Labadze L, et al. A systematic review of educational chatbots: Applications and effectiveness. Computers and Education: Artificial Intelligence. [accessed July 20, 2025] Available at: https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-023-00426-1?utm_source=chatgpt.com

[6] Szigriszt Pazos F. Predictive systems for written message readability: perspicuity formula [Internet]. 1993 [accessed July 20, 2025]. Available at: https://webs.ucm.es/BUCM/tesis//19911996/S/3/S3019601.pdf

[7] Bakhshinategh B, Zaiane OR, ElAtia S, Ipperciel D. Educational data mining applications and tasks: A survey of the last 10 years. Educ Inf Technol [Internet]. 2018 [cited July 20, 2025];23(1):537–553. Available from: https://doi.org/10.1007/s10639-017-9616-z

[8] Belcic I, Stryker C. What are model parameters? [Internet]. IBM; 2024 [cited 2025 Jul 27]. Available at: https://www.ibm.com/mx-es/think/topics/model-parameters

[9] Barrio I. The Inflesz program [Internet]. Legibilidad.com: a website on the analysis of the readability of texts written in Spanish; Jan. 11, 2015 [cited 2025 Jul. 20]. Available at: https://inflesz.com/

[10] Misischia CV, Poecze F, Strauss C. Chatbots in customer service: Their relevance and impact on service quality. In: Proceedings of the 13th International Conference on Ambient Systems, Networks and Technologies (ANT); 2022 Mar 22–25; Porto, Portugal. Procedia Computer Science. 2022;201:1177–1184. Available at: https://www.sciencedirect.com/science/article/pii/S1877050922004689

[11] Fayyad U, Stolorz P. Data mining and KDD: Promise and challenges. Future Gener Comput Syst [Internet]. 1997 [accessed July 20, 2025];13(2–3):99–115. Available

[12] Brandtzaeg PB, Folstad A. Why people use chatbots: authors' version. SINTEF [Internet]. 2017 [cited 20 July 2025]. Available from: https://sintef.brage.unit.no/sintef-

xmlui/bitstream/handle/11250/2468333/Brandtzaeg_Fol stad_why+people+use+chatbots_authors+version.pdf
[13] Brown T. Language models are few-shot learners. NeurIPS [Internet]. 2020 [accessed July 20, 2025]. Available at: https://proceedings.neurips.cc/paper/2020/file/1457c0d6 bfcb4967418bfb8ac142f64a-Paper.pdf
[14] Hugo SE. The Probabilistic Relevance Framework: BM25 and Beyond [Internet]. 2009 [accessed July 20 July 2025]. Available at: https://www.researchgate.net/publication/220613776_T he_Probabilistic_Relevance_Framework_BM25_and_B eyond
[15] IBM. Chatbots [Internet]. IBM; 2023 [accessed July 20, 2025]. Available at: https://www.ibm.com/mx-es/topics/chatbots

**Contribution Roles Table**

| Role | Author(s) |
|---|---|
| Project Management | Luis José Muñiz Rascado |
| Resources | Eugenio César Velázquez Santana |
| Software | Carlos Pérez Calderon (same) Luis Ángel Vázquez Carrillo (same) |
| Validation | Noé Alejandro Castro-Sánchez |
| Writing - Preparation of the original draft | Carlos Pérez Calderon (same) Luis Ángel Vázquez Carrillo (same) |

This work is licensed under an international Creative Commons Attribution 4.0.