The National Higher School of Artificial Intelligence

Natural Language Processing Course Project

Fall 2024

Dhakirate-Al-Djazair:

Reviving Algerian History Education with Large Language Models and Retrieval-Augmented Generation

ARAB Sarra

Abstract

This paper discusses "Dhakirate-Al-Djazair," an original application that will change how the history of Algeria is taught, using large language models and retrieval-augmented generation. History in Algeria spans several epochs and levels of education, none of which is properly supported with traditional resources. In this respect, our project integrates natural language processing and deep learning methodologies in order to extract and process primary texts from PDF sources, curate a corpus organized by both historical era and educational stage, and generate interactive gamified exercises. By providing the content tailored for primary, junior, secondary, and university students, the system will contribute to deepening historical understanding and increasing overall engagement. Our approach points to the potential of NLP-driven educational tools in bridging knowledge gaps and fostering contextualized learning, thus empowering students with an enriched grasp of Algeria's diverse historical heritage.

Keywords : Algerian History; Large Language Models; Retrieval-Augmented Generation; Adaptive Learning; Educational NLP.

Introduction (1 page to 1.5 pages)

From ancient influences to modern-day developments, Algeria's history spans an amazing breadth of eras and events that shaped the nation's identity. Teaching such a wide range of material to students of varying educational levels presents considerable challenges. Traditional classroom approaches and resources sometimes struggle to capture the richness of historical narratives or provide the interactive engagement necessary to spark students' curiosity. In addition, access to and organization of relevant source materials, especially those targeted at different key stages, is often inadequate. As a result, the students may miss the multi-layering complexity of each historical period and fail to grasp the full depth of the cultural and political forces that shaped contemporary Algeria.

It is in this context that the "Dhakirate-Al-Djazair" project embarks on the exploitation of new technologies, especially large language models and retrieval-augmented generation, to offer an adaptive holistic learning environment for Algerian history. It will extract and process textual information from PDFs, textbooks, and online materials to create a well-curated corpus of historical texts that is aligned with particular eras-e.g., Pre-Islamic, Ottoman Empire, War of Independence-and educational levels: primary through university. It goes beyond the mere aggregation of information and aspires to deliver content in a manner that would appeal to today's learners, incorporating NLP-driven techniques for text classification and question-answer generation.

Most importantly, the project underlines how digital innovations can enrich and democratize learning experiences. Other than static, ready-made textbooks, "Dhakirate-Al-Djazair" uses advances in machine learning to adapt the content to learner needs-so primary-school learners will only see explanations in a simpler manner, while older students and university scholars could delve deeper into analyses. Besides this, the platform provides for gamification to promote better comprehension and tracking of the learner's progress to motivate him for continued practice. This methodology not only tackles pragmatic problems in teaching such complex historical subjects but also leads one step into the future of pedagogy, where technologies empower students to navigate intricate information ecosystems.

In essence, the "Dhakirate-Al-Djazair" will be much more than just a technological tool; it will be a bridge that connects students with the tapestry of Algeria's history in accessible and dynamic ways. The proposed project would develop state-of-the-art NLP capabilities in a structured manner for content curation, which might lead to improved academic performances but also more profound cultural appreciation. This paper now examines certain aspects of the technical components, data collection strategies, and methods of evaluation, together with the projected outcomes as a basis for a new paradigm in history education.

State of the Art

2.1 Evolution of NLP and Large Language Models

Natural Language Processing (NLP) has rapidly evolved over the past decade, and large language models have been major factors in that change. Early work in NLP was based on rule-based systems, then shifted toward statistical and machine learning methods. More recently, the publication of transformer-based models like BERT and GPT has truly upended the landscape. These models are typically trained on massive amounts of text data and exhibit remarkable capabilities in natural language understanding, generation, and transformation. In education, such LLMs have been particularly effective at text summarization, question-answering, and machine translation-all of which are valuable skill-sets in the context of history. For instance, BERT implemented bidirectional contextualization and gave a critical fillip to reading comprehension tasks, where deeper insights from the passages were realized. While GPT-based models went one step further and showed surprising generative properties, being able to generate coherent text segments, respond relevantly to given prompts, and even adapt to topics. It is this generative strength that has subsequently opened ways for question-answering systems that can tackle complex queries with context-sensitive responses.

2.2 Retrieval-Augmented Generation for Domain-Specific Knowledge

While large language models are very capable with respect to general language tasks, purely generative models often tend to make factual inaccuracies, especially in domains that are specialized or lesser-known. In response, researchers and practitioners have explored "retrieval-augmented generation", an approach which integrates external data sources into the generation process. Models such as RAG, which couples a neural retriever-to fetch relevant passages from a text corpus-with a generative model, to formulate answers, ground responses in vetted data.

Retrieval augmentation is interesting in a completely different light when it concerns history education: drawing on true facts, timelines, and nuanced interpretations that curated texts have guaranteed accurate and with rich contexts, by linking generated outputs to specific references from the text corpus, it enhances traceability and therefore transparency for a developer dealing with subject matter as sensitive and wide-ranging as the history of Algeria. Recent research in this area, such as open-domain question-answering, has emphasized the importance of retrieval methods for the reduction of hallucinations, improvement of factual consistency, and for enabling updates or extensions of the knowledge base without full model retraining.

2.3 Current Educational Tools and Their Shortcomings

A number of educational platforms already use NLP to some extent. Online tools such as Duolingo or Babbel use NLP-driven exercises for language learning, while Khan Academy and Coursera provide personalized learning experiences, though not typically anchored in advanced LLMs. In history education, a few systems offer digitized textbooks, interactive quizzes, or historical simulations; however, these often rely on static materials or simplistic multiple-choice question banks.

A common shortfall with existing tools is a lack of domain specificity. Generic question-answering systems, for instance, may provide general overviews and fail to address the specific contexts or particular historical narratives that are taught in school. Another recurring limitation is related to a lack of mechanisms for robust differentiation by learner level. Most educational systems do not dynamically adapt the reading complexity or scope for primary versus university-level students, thus limiting their utility across diverse age ranges. Historical texts, in particular, introduce unique classification, translation, and context retention challenges for works that span several centuries and multiple languages. Many current general-purpose NLP solutions are not optimized to such domain-specific complexity.

2.4 Gamification and Adaptive Learning in Historical Education

Gamification-introducing game-like elements such as points, badges, or quests-has been one of the main approaches aimed at improving the learners' engagement. Well-designed gamified exercises can encourage more sustained motivation and provide for effective progress tracking while fostering better knowledge retention. Applications like Kahoot! are but one example of how quizzes and playful competition is changing the learning experience, but such options may not bring in deep personalization or advanced NLP techniques. These types of systems depend on human-created quizzes that do not adapt to question difficulty either based on learner profile or on real-time performance.

Recent academic work in adaptive learning systems has shown the potential of combining NLP-based text difficulty classifiers with real-time analytics on learner performance. In such systems, reading material and quizzes could be tailored to a user's proficiency for a more personalized experience. Though still emerging in mainstream products, adaptive learning research underlines the value of LLMs for generating and grading open-ended responses, performing rapid summarization of content, and identifying misconceptions through free-text analysis.

2.5 Novelty and Positioning of "Dhakirate-Al-Djazair"

Considering the above progress, there is still a gap in specialized historical domains-especially regions

or time periods that are not well-represented in large public datasets. The history of Algeria presents examples of such history: pre-Islamic influences, Ottoman governance, French colonization, and the War of Independence create a holistic demand for culturally and contextually profound narrations. These models of generic question-answering usually result in a lack of verifiable information or gross simplification on such specific matters.

“Dhakirate-Al-Djazair" tries to fill this gap by combining the power of retrieval-augmented LLMs with a custom-developed corpus sourced from relevant textbooks, PDFs, and academic articles. Materials are categorized according to historical era and educational stage in classification so that learners may get only what they need. Further, for interactive exercises that include real-time feedback, gamification principles have not been overlooked in designing the system, not limited to just text-based question-answering. While each component separately has emerged-LLMs, retrieval augmentation, adaptive learning, gamification-in the research community, their holistic combination specifically for education in the history of Algeria is relatively unexplored.

The novelty of the solution thus consists in the duality of focus: first, on domain specificity at a granular level (Algerian history across school levels) and second, on using state-of-the-art NLP methodologies, namely retrieval augmentation and language-model fine-tuning. Merging these two sides, the project seeks to provide a specifically tailored, complete, and interactive platform that respects factual accuracy while bringing added value for students in the understanding of their cultural heritage.

All work reviewed here together supports the case for how large language models and retrieval-augmented approaches have the potential to dramatically shift education practices. Each section of the following report outlines in detail how "Dhakirate-Al-Djazair" applies these very advancements to offer a new paradigm for history and beyond in innovative, adaptive, and contextually robust learning.

Data Set Building

The project will design a quiz generation system and a chatbot on the education of Algerian history, taking data as the central element. To this end, for effective development of these functionalities, we have created three different datasets, each designed to handle different aspects of Algerian history, namely:

- Algerian History Educational Dataset: This provides comprehensive information about the historical milestones of Algeria, covering key events, eras, and aspects of culture.
- Date-Event Dataset: This contains major historical events in Algeria and the dates on which they occurred.
- Personality Description Dataset: This outlines all the key historical personalities of Algeria, along with their contributions and influence on the overall development of the nation.

Both these datasets are designed to cater to the demands of different quiz question types while providing conversational assistance through the chatbot for a better learning experience.

1. Data Collection

Methods Used to Gather the Dataset:

- Manual Collection: We manually collected content from educational PDFs, textbooks, and historical resources, including teacher lectures, academic summaries, and research articles. These resources were crucial in creating a broad overview of Algerian history.
- Sources: The primary sources for the dataset included Ministry-approved textbooks, digital archives, historical websites, and academic publications. These sources provided authentic and reliable content for the dataset.
- Challenges: The volume of historical content was huge, and extracting relevant and accurate material from such diverse formats was one of the major challenges in data collection. In this regard, we have collected, tagged, and organized the content in a very systematic way. Another challenge was handling non-Arabic texts; we leveraged translation APIs to make the content available in Arabic.

1.1 Algerian History Educational Dataset:

Initial Process: The educational content was manually collected by copying and pasting from lectures, academic summaries, and research articles. Collected data were categorized into different educational stages, historical topics, and historical eras in order to fit the curriculum.

Manual Tagging: Each piece of content was tagged with appropriate fields so that it would be easy to retrieve and organize.

Fine-tuning the chatbot: We digitized more historical content from scanned books through OCR, thus converting pages into digital texts. These were then processed to enrich the dataset with more minute details

related to history.

Model training: A structured dataset and a book-based dataset were integrated to form an all-inclusive base that the rage model used for training.

1.2 Date-Event Dataset:

The extraction process of PDF contents into usable data for the Date-Event Dataset began by extracting relevant sections from educational PDFs. Since some of these PDFs were not directly extractable, screenshots of pages containing dates and events from history were captured to capture what was needed.

This approach allowed us visually to capture that essential content, which text-based extraction was impossible to do. These were subsequently uploaded on Google Lens, which effectively transcribed the image of text into machine-readable, editable text. It was instrumental to the conversion of the visual into a form suitable for processing and analysis. Manually cleaned extracted text for accuracy and consistency: checking dates and events to see that they fit the historical context in which they had taken place. Manual intervention was required to ensure the data were reliable and consistent.

Augmentation with Existing Content:

While collecting these for the first time, we felt there was a serious lack of sufficient date-event data. In order to handle this, we went back to the previously created Algerian History Educational Dataset and extracted further dates and events.

1.3 Personality Description Dataset

Similar to the Date-Event Dataset, much data in the Personality Description Dataset came from PDFs.

3. Annotations and Preprocessing:

The annotations have been used to structure and tag the content for easy retrieval.

Fields Explanation:

- EducationalStage: Specifies which educational stage this content is appropriate for.
    - PS: Primary School
    - JS: Junior School
    - HS: High School - General
    - HSS: High School - Scientific
    - HSL: High School - Literature
    - UNI: University

- HistoricalEra: Represents the period to which this content belongs. For example:

- Prehistoric Era (العصور ما قبل التاريخ)
- Ancient Era (العصر القديم)
- Byzantine Era (العصر البيزنطي)
- Islamic Era (العصر الإسلامي)
- Ottoman Era (العصر العثماني)
- Colonial Era (عصر الاستعمار)
- War of Independence (حرب الاستقلال)
- Post-Independence Era (فترة ما بعد الاستقلال)
- Modern Era (العصر الحديث)

It tags the historical context to which the content refers.

- Topic:
  - A concise description of the subject covered, such as "French Colonization" or "Algerian Resistance.".
  - This field provides a short heading or title for the historical content.
- Content:
  - A detailed description of the historical topic. It includes the main text related to the event, figures, or periods under discussion.
- Source:
  - Indicate the source of the content. Examples include:
    - Original: The content is original creation from an academic source, lecture, or research.
    - Reference(s) to be provided for specific references sourced from textbooks, research articles, etc.
- Level:
- Reflects the level of the content with regard to educational purpose within the curriculum; for example, Level 1 = beginner, Level 2 = intermediate, Level 3 = advanced.

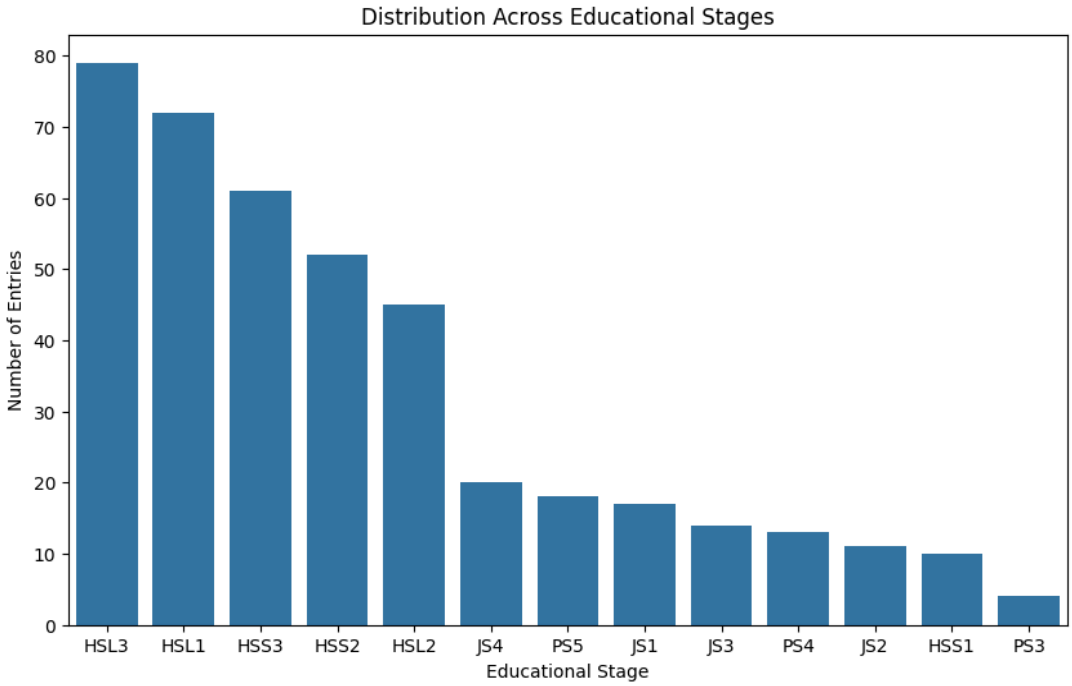
4. Dataset Characteristics

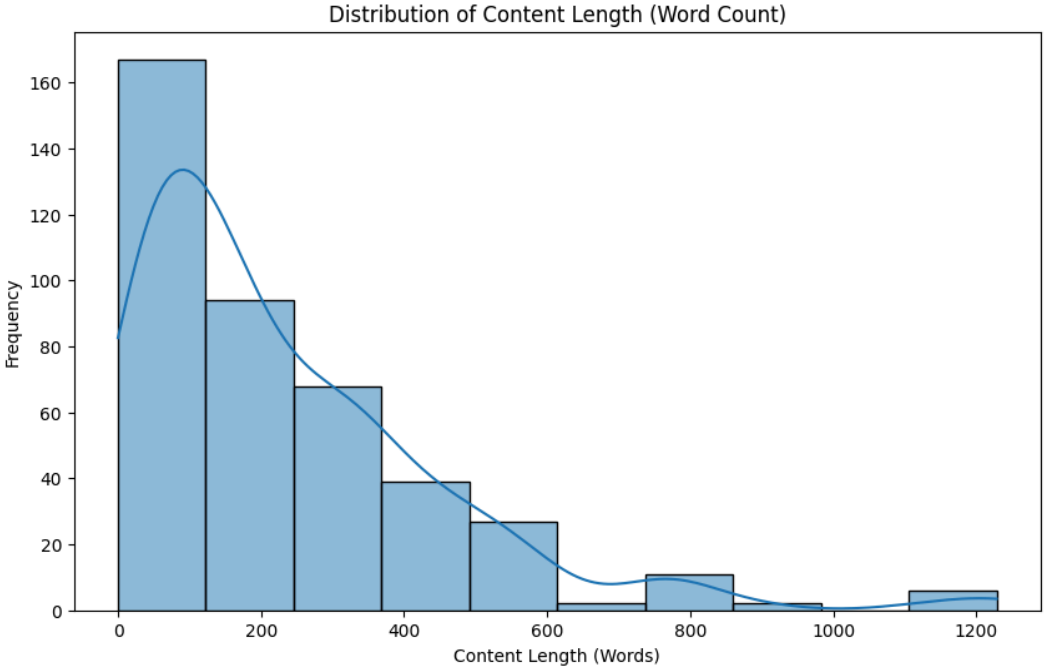
Digitized Books Dataset (17 Books):

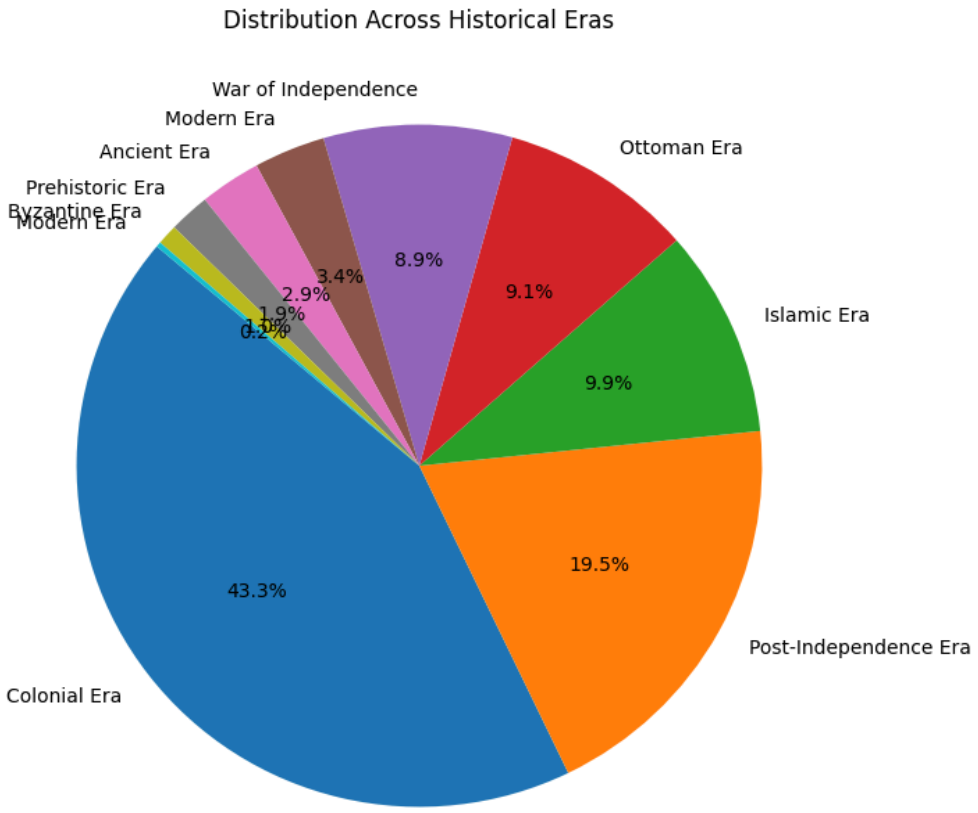
- High School Scientific (1, 2): 2 books
- High School Literature (1, 2): 2 books
- Junior School (1, 2, 3, 4): 4 books
- Primary School (3, 4, 5): 3 books
- University: 9 books

Content for Quiz:

- Total Entries: The dataset contains 416 entries in total.
- Total Word Count: The total word count of all entries in the dataset is 97,734 words.
- Average Length of Entries: The average length of an entry is 234.94 words.
- Number of Topics: There are 256 unique topics across all entries in the dataset.




Distribution of Content Length (Word Count)
Frequency
Content Length (Words)
0
20
40
60
80
100
120
140
160
0
200
400
600
800
1000
1200


Distribution Across Historical Eras
War of Independence
Modern Era
Ancient Era
Prehistoric Era
Byzantine Era
Modern Era
Ottoman Era
Islamic Era
Post-Independence Era
Colonial Era
8.9%
9.1%
9.9%
19.5%
43.3%
3.4%
2.9%
1.9%

System Design

The project aimed to develop an educational platform for Algerian students in order to learn about the history of Algeria through gamified and interactive methods. We used **Large Language Models** and **Retrieval-Augmented Generation** in order to continuously generate relevant information for the users. One of the big challenges was how to create this platform with existing data resources, considering the interaction of the platform and the students will be in an attractive and amusing manner that could help the process of learning.

1. Platform Architecture:

We designed this platform to cater to all educational levels of Algerian students and provide multiple ways of interaction to ensure that information effectively reaches the students. The platform consists of four main sections:

1. **Quiz Section:**
   This section is dedicated to quizzes where students can engage in various types of assessments, including multiple-choice questions, date/event completion, and personality cards, allowing them to test their knowledge in a fun and interactive way.
2. **Lesson Section:**
   Lessons are presented interactively, allowing students to engage actively with the content. Students can save specific information for future reference and even chat with *Dakira*, the chatbot, about the lesson topics for deeper understanding.
3. **Progress Section:**
   This section displays the student's progress, showcasing statistics related to their performance in quizzes and learning levels, giving them insight into their growth and areas to improve.
4. **Dakira Chatbot Section:**
   In this section, students can interact with *Dakira*, the chatbot, to ask questions or explore any topic related to Algerian history, creating a more personalized and interactive learning experience.

Additionally, the platform will include a student authentication system to ensure the secure handling of personal information, allowing for a more tailored learning journey based on individual progress and preferences

2. Design The Architecture:

2.1 Quizzes: Three distinct quiz types were designed and implemented as part of our platform to foster variety and enhance interactivity in the student learning experience. Each quiz is developed with a unique structure to cater to different learning styles and ensure a dynamic educational environment:

- **2.1.1 Quiz1:** The first type of quiz implemented in the platform is a multiple-choice question format, where students are exposed to a question with several possible answers and must then choose the right one. The design of the quiz considers the student's educational stage and level, whereby the questions must be aligned with the academic progress that corresponds to him. Another important aspect of this

method is the inclusion of the **level of progress** by the student. For maximum learning to occur, there is a need to reduce the number of questions that have already been answered and hence the need for the exposure of the student to new questions that will help in improving the learning process.

Technically, this type of quiz can be developed using either of two methods:

LLM-based approach: A large language model (LLM) is utilized to generate questions and corresponding answer options based on relevant documents tailored to the student's educational stage and level. To prevent the recurrence of previously encountered questions and ensure the delivery of novel content, we have implemented a filtering mechanism. This filter is designed to minimize the number of documents that have already been used to generate questions previously solved by the student. The filtering process operates by calculating the **cosine similarity** between the embeddings of the retrieved documents (content) and the embeddings of previously answered questions which represent **the level of progress** of the student. The cosine similarity score quantifies the degree of similarity between the content and the solved question. Only those documents whose similarity score falls below a predefined threshold are considered for use by the LLM to generate new questions. This threshold is determined through experimentation and testing, optimizing the separation between new and previously encountered content.

By applying this filtering technique, the system ensures that only new, unsolved documents are selected for question generation, preventing redundancy and ensuring the novelty of the content. This approach guarantees that the content is both relevant to the student's current progress and aligned with their educational level, enhancing the overall learning experience.

Question Bank Approach: The system will utilize a pre-built question bank, which is continuously expanded over time to maintain variety and ensure the inclusion of relevant questions. As new questions are added, they are selected based on the student's educational year and level, ensuring alignment with their current academic stage. To further enhance the novelty and relevance of the questions, a filtering mechanism is employed. This filter ensures that only new, unsolved questions are passed for use in generating queries. These selected questions are then used to extract relevant documents from the structured dataset through a Retrieval-Augmented Generation (RAG) framework. The system first retrieves the most pertinent documents based on the question, which serve as the source of content for the LLM. Both the question and the retrieved documents are subsequently fed into the LLM, where the documents play a crucial role in response extraction. The LLM is then responsible for ensuring the accuracy and correctness of the generated answers by synthesizing information from the retrieved documents, thus ensuring a high-quality and reliable output.

We chose to adopt the first approach due to its simplicity and lower computational requirements.

- **2.1.2 Quiz2:** The second type of quiz is designed for dates or events, where the student is required to identify the corresponding event or date, respectively. This format is particularly beneficial as it encourages not only the memorization of dates and events, but also the understanding of their

chronological relationships through presenting them in their chronological timeline order

For the implementation of this quiz, we used dates_events dataset. In line with the approach used in the first type of quiz, we leverage the student's educational stage, level, and progress (which tracks previously solved dates and events) as a filtering mechanism implemented to eliminate events that have already been solved by the student to ensure appropriate question selection. Additionally, we standardize all dates in the dataset to ensure accurate retrieval of corresponding events and correct evaluation of student responses. When the student provides an answer related to an event, we employ embeddings and cosine similarity metrics to assess the similarity between the student's response and the correct event corresponding to the given date. This allows the system to accommodate minor discrepancies, such as misspellings or omitted letters, by calculating the similarity between the student's input and the correct event. A predefined similarity threshold is applied to determine whether the answer is correct, ensuring that small variations in input do not hinder the evaluation.This threshold set after a series of tests and experiments

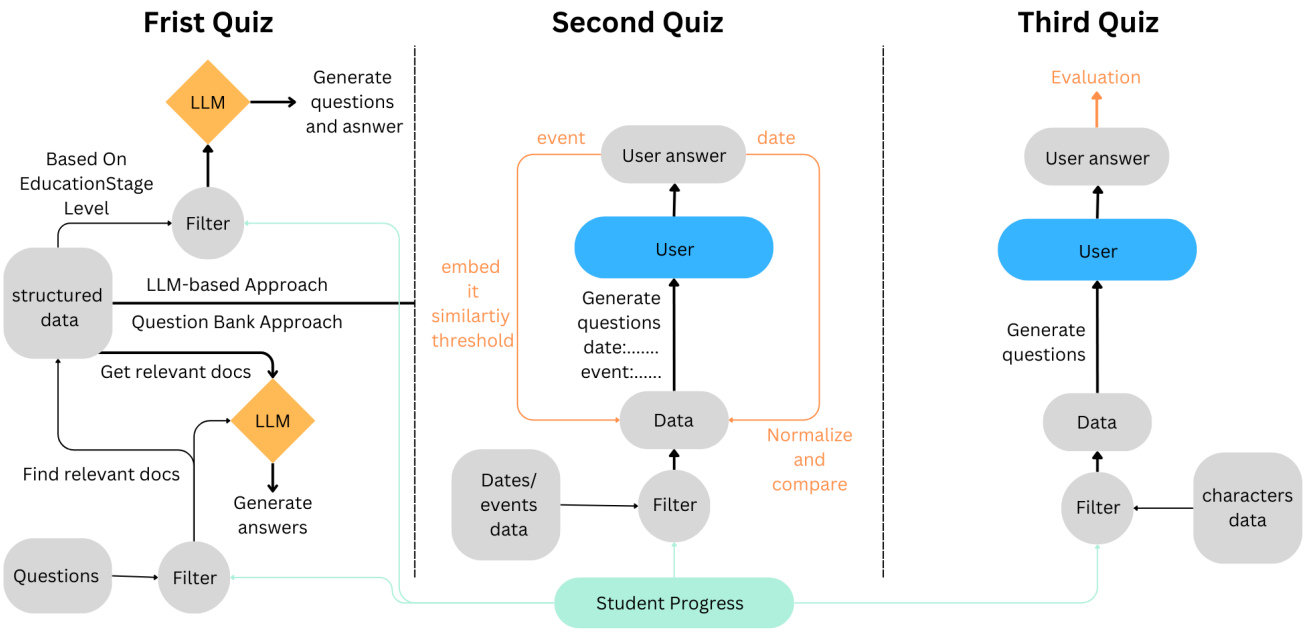
Similarly, when the student is tasked with identifying the correct date for a given event, we apply natural language processing (NLP) techniques(digits handling, regular expression..) to normalize the student's input, facilitating a comparison between the student's answer and the correct date. This normalization process ensures that minor differences in formatting or phrasing do not affect the accuracy of the evaluation.

- **2.1.3 Quiz3:** The third type of quiz involves providing a list of historical personalities along with their corresponding descriptions, where the student is tasked with matching each personality to the correct description. This format is particularly effective in promoting the association of historical figures with their key achievements or roles, thereby fostering a more profound understanding of their significance in history.

  For the implementation of this quiz, we used specialized dataset that links historical personalities to their respective descriptions. Consistent with the other quiz types, the student's educational stage, level, and progress are taken into account to ensure that the questions align with the student's current knowledge and learning path. In this quiz, students interact with the system by either dragging and dropping or selecting the correct matches between personalities and their descriptions from randomized list. The randomization of options is designed to enhance the challenge and engagement of the quiz, encouraging the students to think critically and make accurate associations.

  The system performs real-time evaluation of the matches, verifying the correctness of each pair as the student progresses through the quiz. Additionally, the student's progress is tracked to prevent the repetition of previously solved personalities and descriptions, ensuring a systematic progression through new content. This mechanism promotes a deeper learning experience by continually introducing fresh material tailored to the student's educational stage and level.

This figure below show the entire quizzes architecture:



2.2 Chatbot: The platform will feature a domain-specific chatbot focused on Algerian history, enabling students to ask questions and receive answers closely aligned with their academic program. This will be achieved by utilizing RAG with our curated dataset to define the domain knowledge. By integrating the retrieved knowledge with the user's query, we will provide the LLM with the necessary context to generate answers that are both relevant to the student's question and aligned with the retrieved documents. We call this chatbot **Dakira** and it will be introduced as follows:

- **Topic-based Chatbot:** The platform incorporates a topic-based chatbot to facilitate interactive learning and enhance student engagement with educational content. Students can review topics (lessons) directly retrieved from the platform's structured database. To further enrich this experience, we plan to integrate a large language model (LLM) in future iterations, enabling the generation of additional contextual information for specific topics. Each topic page will include an option for students to engage in focused discussions with the chatbot, Dakira. The chatbot's responses will be context-specific and tailored to the content of the selected topic. This design ensures that interactions remain relevant and aligned with the student's query.

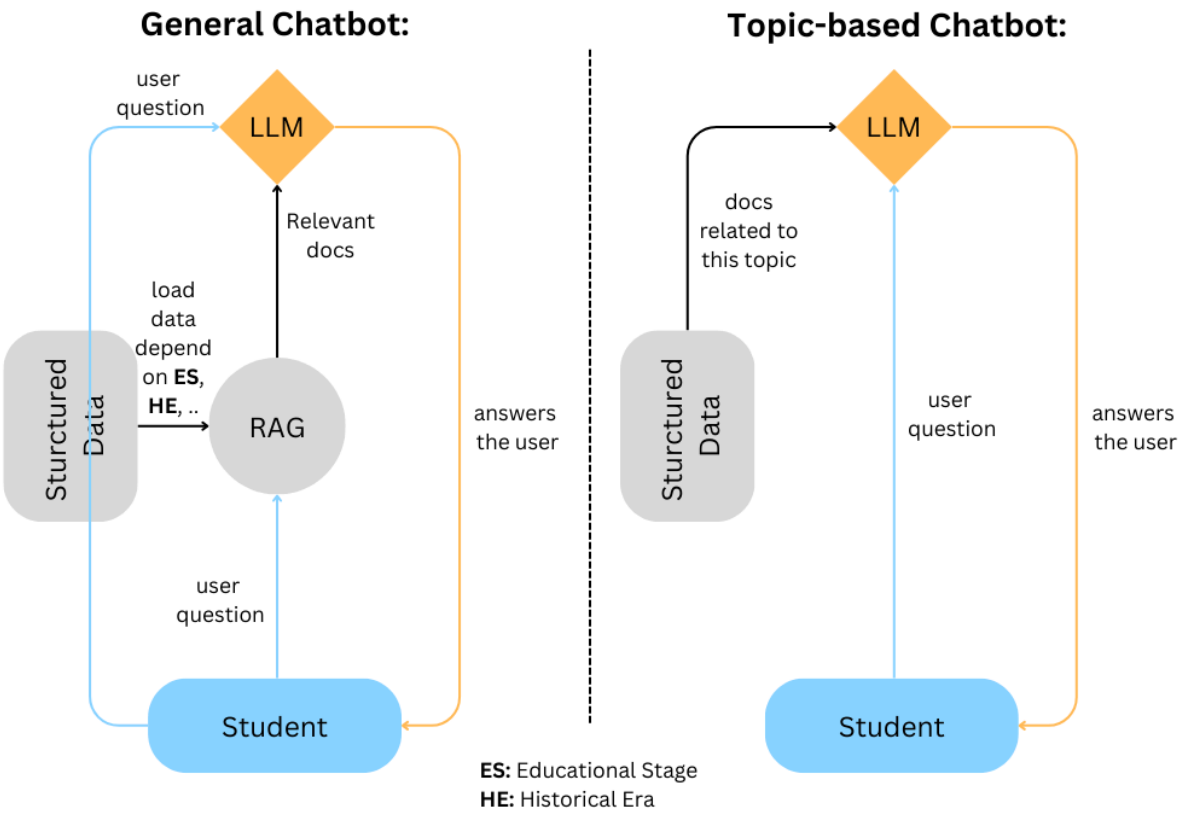
  To implement this functionality, the system will provide the LLM with only the documents pertinent to the topic under discussion, along with the student's query. To facilitate precise information retrieval, a Topic attribute is included in the dataset, enabling the identification and extraction of relevant documents. By limiting the scope of the chatbot's input to topic-specific documents, the system ensures that the generated responses are accurate, contextually appropriate, and educationally relevant.

- **General Chatbot:** The second implementation of the chatbot is designed for general use, where the chatbot's content is no longer restricted to a specific topic. Instead, it utilizes the entire corpus of data to retrieve relevant documents in response to user queries, employing the Retrieval-Augmented Generation (RAG) approach. This implementation is based on a model-driven embedding strategy using the

Alibaba-NLP/gte-multilingual-base model, which leverages L2 distance as the similarity metric for document retrieval.

After extensive experimentation and comparative analysis (detailed in the subsequent section), this model was found to exhibit superior performance in RAG-based retrieval tasks. By integrating this approach, the chatbot ensures a comprehensive and contextually accurate interaction, allowing users to query the system on a broad range of topics while maintaining a high standard of response relevance and precision.

The figure below show the chatbot architecture:



3. Methodology and Development:

Quiz1:

- Data source was mainly from the structured data that contains necessary features such as *Level*, *EducationStage*, and *Content*.
- Clean the data using simple regex to remove redundant new lines `/n`, and it is worth mentioning that we don't want to remove all non-Arabic characters since, when dealing with historical data, we encounter dates like 12/12/1988 or 12-12-1988, so `/` and `-` are important in our dataset. This cleaning is done on the *Content* attribute.
- Calculate the stemmed *Content* using Tahaphyne[1] stemmer to be used in generating embeddings.
- Generate embeddings of all stemmed *Content* rows and save them in the data `.pkl` file.
- Calculate the embeddings of solved questions that are in the student's database block.

- Calculate the cosine similarity and eliminate similar ones.
- Feed the LLM *K* rows of *Content* to generate questions and multiple-choice answers (*K* is a parameter)

```
# Example usage:
educational_stage = "HSS3"
level = 1
num_of_questions = 3
filtered_data = filter_similar_content(data, level, educational_stage, answered_questions_embeddings)
questions_with_options = generate_question_without_repeats(filtered_data, num_of_questions)
```

Quiz2:

- Perform the same process of cleaning and embedding the data for dates and events.
- Embed the solved events and dates from the student database.
- Eliminate the solved events and dates.
- Generate *K* questions with event-based and date-based questions (*K* is a parameter).
- Evaluate the student's answers through similarity evaluation for events and normalization and comparison for dates

```
evalute_answer(questions)
```

```
[0.81315344]
wrong answer
[0.75432069]
wrong answer
correct answer
wrong answer
wrong answer
```

- As observed, the similarity values for event-based student answers were tested across many cases.
- After extensive testing, the appropriate threshold for cosine similarity was determined to be **0.84**

Quiz3:

- The steps for Quiz 3 are nearly identical to the previous quiz on person_dataset.
- Evaluation in this quiz is straightforward, relying on a similarity check.
- Since the quiz is designed as a card-matching game where students link a person's name to its definition, no direct student input is required.

Chatbot:

- For the Topic-Based Chatbot, the data used was the same as the data used for Quiz 1, and the cleaning and preprocessing steps were identical.
- For the General Chatbot, the data used was larger, resulting from merging the structured data with the book data.

- The embedding was calculated only for the General Chatbot's content rows. For the Topic-Based Chatbot, the relevant documents are those that share the same topic, so there was no need to use RAG.
- The user query is processed through the same preprocessing steps and embedding model used for Quiz 1.
- Using FAISS index, the most similar K documents (where each document corresponds to one row in the dataframe) are retrieved.
- For the LLM choice, Google Gemini 2 9B it was selected due to its strong performance in generating Arabic context and answering historical questions.
- The platform hosting the model is Qroq, which enables API requests to have a very short memory, where each API request is cleaned in the subsequent request.
- A data structure called "history" was created to store the user's question and the chatbot's answer for a specific session, the user is able to ask related questions based on previous queries, allowing for more coherent and contextually relevant interactions.By maximizing the context length this ensures that the LLM can handle the three components: the student's query, relevant documents, and the history.

Results and Analysis:

To ensure the development of a high-performance system, it was necessary to evaluate the various components of our project and identify key starting points. As discussed in the previous section, the project focuses on creating a platform to educate students about the history of Algeria through gamification, incorporating interactive quizzes and a chatbot. The system is built upon well-structured data embeddings, the application of Retrieval-Augmented Generation (RAG) techniques, and the accuracy of large language models (LLMs). This section provides a detailed analysis, including a comparison of RAG techniques, development results, and their evaluation:

RAG Comparison and Analysis:

The tests and comparisons were conducted primarily on the **Structured Dataset**, where the target text was located in the Content column. Various embedding techniques, including **TF-IDF**, **BM25**, **Word2Vec**, and embeddings based on **transformer models**, were tested and evaluated using different metrics. These metrics include:

- **Precision** *Retrieved Relevant Documents / Top k Retrieved Documents*
- **Recall** *Retrieved Relevant Documents / Total Relevant Documents*
- **Mean Reciprocal Rank (MRR)** *1 / Rank of the First Relevant Document*
- **Cosine Similarity (Average)**
- **Normalized Discounted Cumulative Gain (nDCG)** *DCG / IDCG*
  Here, DCG refers to the Discounted Cumulative Gain, and IDCG refers to the Ideal DCG.
- **Mean Average Precision (MAP)**

It is worth mentioning that each RAG technique was tested on four different versions of the Content corpus, which are: **Original Content**, **Stemmed Content** (using predefined stemming rules), **Tashaphyne Stemmed** Content (processed using the Tashaphyne stemmer), **Lemmatized Content** (using Madamira System)

Additionally, the queries were generated for each document by randomly sampling a set of continuous words with varying lengths **(5**, **10, 20** words) and selecting them from different positions within the text (start, middle, end)

Word Count Methods: We utilized TF-IDF and BM25 on the four different versions of the Content corpus, and the results are presented as follows**:** a Figure**.4** for TF-IDF, Figure**.5** for BM25 and Figure.**6** for both:

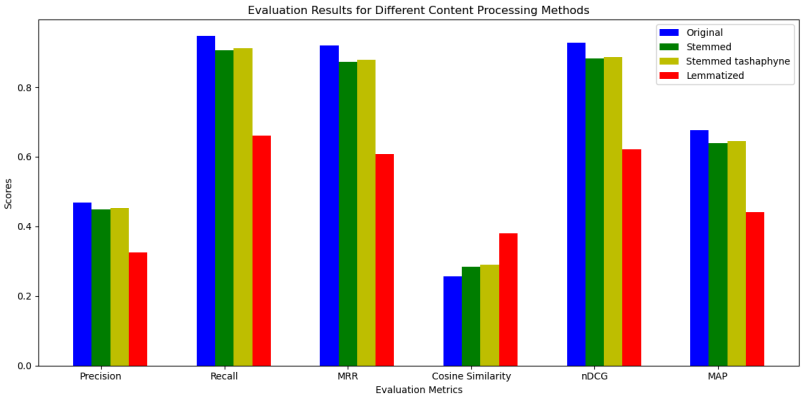


Figure.4: TF-IDF results on different versions of the Content

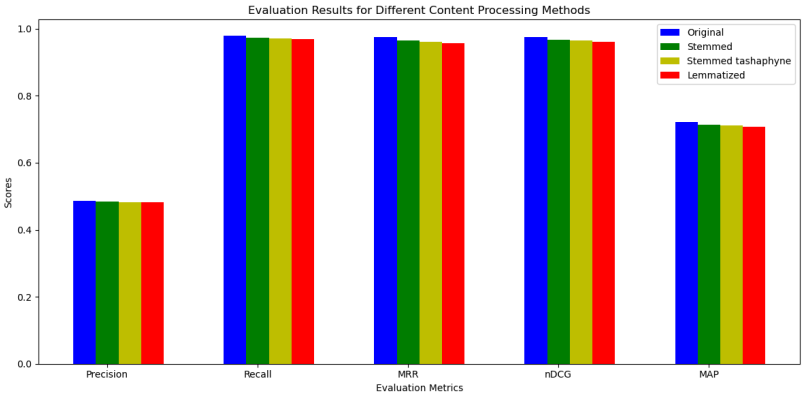


Figure.5: BM25 results on different versions of the Content

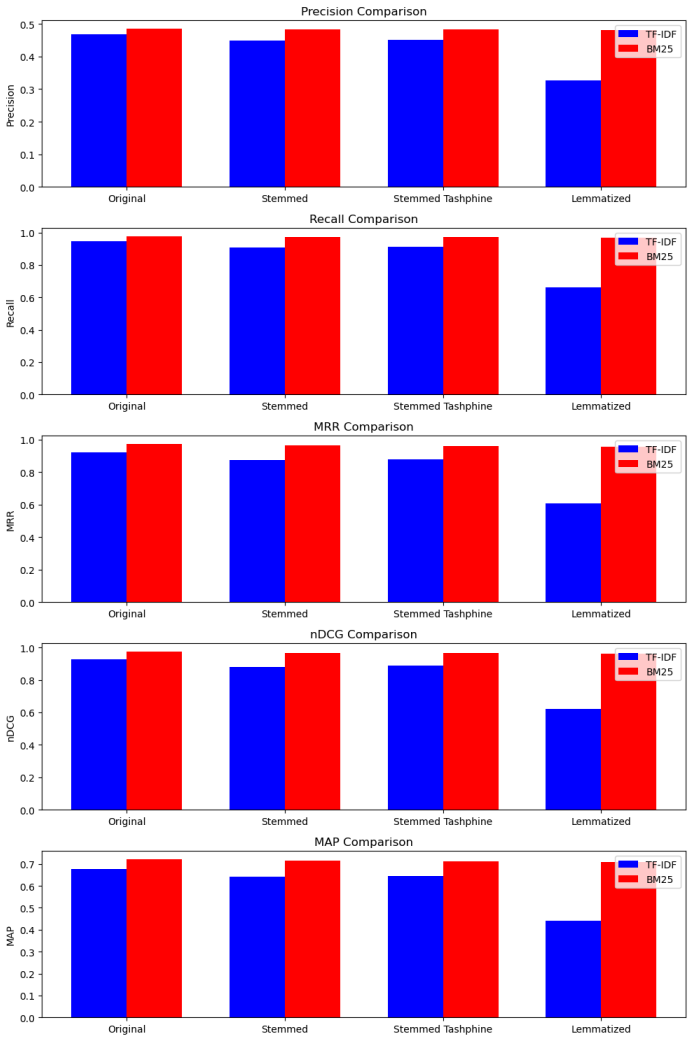


Figure.6: Comparison of TF-IDF and BM25

[It is important to note that the values of Recall are generally high since, in most cases, each query is associated with only one or two relevant documents, as shown in the distribution of the number of relevant documents per query in Figure.7 While the Recall values are often high, they still indicate that the technique performs well, considering that the majority of queries return only one or two relevant documents. Similarly, Precision does not exceed 50%, largely due to the retrieval limit of three documents per query. Since most queries have one or two relevant documents, the Precision remains under 2/3, which should still be regarded as good performance. These considerations also support the performance observed in MAP.]

Content Type	Precision	Recall	MRR	nDCG	MAP
TF-IDF					
Original Content	0.4678	0.9464	0.9208	0.9273	0.6763
Stemmed Content	0.4485	0.9068	0.8733	0.8816	0.6399

Stemmed Tashaphyne Content	0.4520	0.9123	0.8782	0.8868	0.6452
Lemmatized Content	0.3260	0.6613	0.6087	0.6208	0.4414
BM25					
Original Content	**0.4862**	**0.9784**	**0.9745**	**0.9762**	**0.7222**
Stemmed Content	0.4834	0.9738	0.9641	0.9673	0.7134
Stemmed Tashaphyne Content	0.4833	0.9722	0.9616	0.9650	0.7107
Lemmatized Content	0.4818	0.9688	0.9580	0.9608	0.7070

The analysis of the TF-IDF results reveals that **original content** outperforms all other preprocessing methods, achieving the highest **MRR (0.9208)**, **MAP (0.6763)**, demonstrating superior document ranking and retrieval performance. **Stemmed content** shows a moderate decline,, but still maintains solid retrieval performance. **Stemmed Tashaphyne content** performs similarly to stemmed content. In contrast, **lemmatized content** shows the worst results, with a significant drop in both **MRR (0.6087)**, **MAP (0.4414)**, indicating that lemmatization negatively impacts the system's ability to rank relevant documents. Thus, **original content** remains the most effective preprocessing method, while **lemmatized content** is the least suitable. BM25 outperformed TF-IDF across all evaluation metrics, achieving the highest performance with the original content. Specifically, BM25 achieved a peak values of **0.9784** for Mean Reciprocal Rank (MRR) and of **0.7222** for MAP, surpassing the results from TF-IDF. The stemmed and Tashaphyne-stemmed content also demonstrated strong performance, though slightly trailing behind the original content. Furthermore, a notable improvement was observed in the lemmatized content when compared to the TF-IDF results, indicating a closer alignment to the performance of other ranking systems. These findings underscore the **superior performance** of BM25 over TF-IDF, particularly with the original dataset

Here some queries and retrieved documents using TF-IDF and BM25:

- **TF-IDF on Original Content:**

```
query = "الهجمات الصليبية"
search(query,data,tfidf_vectorizer, tfidf_matrix)
```

Python

```
الهجمات الصليبية
Topic: درس حول الأسطول والبحرية الجزائريةللسنة الثالثة متوسط
Content: على اعتناق المسيحية الإتاوات مبالغ مالية تدفعها السفن البحرية الأوربية العابرة للبحر الأبيض المتوسط مقابل حماية الأسطول البحري الجزائري لها
Cosine Similarity: 0.1071

Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
Cosine Similarity: 0.0930

Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
Cosine Similarity: 0.0930
```

- **BM25 on Original Content:**

```
query = "الهجمات الصليبية"
search_bm25(query, data, bm25)
```

Python

```
Processed Query: الهجمات الصليبية
Topic: درس حول الأسطول والبحرية الجزائريةللسنة الثالثة متوسط
Content: على اعتناق المسيحية الإتاوات مبالغ مالية تدفعها السفن البحرية الأوربية العابرة للبحر الأبيض المتوسط مقابل حماية الأسطول البحري الجزائري لها
BM25 Score: 8.0426

Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
BM25 Score: 7.2054

Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
BM25 Score: 7.2054
```

Both methods retrieved the same documents with identical rankings. Upon review, the retrieved documents were found to be relevant and contained information pertinent to the query context. In terms of precision and accuracy, the original content generated by both methods was nearly identical

- **TF-IDF on Lemmatized Content:**

```
query = "الهجمات الصليبية"
search(query, data, tfidf_vectorizer_lm, tfidf_matrix_lm, lemma=True)
```

Python

```
2025-01-06 14:04:25,004 - INFO - Processing 1 texts in 1 batches
Lemmatizing:   0%|          | 0/1 [00:00<?, ?it/s]2025-01-06 14:04:25,006 - INFO - Starting MADAMIRA server...
2025-01-06 14:04:35,009 - INFO - MADAMIRA server started successfully
2025-01-06 14:04:35,010 - INFO - Input XML file generated successfully: input_0.xml
2025-01-06 14:04:35,887 - INFO - Stopping MADAMIRA server...
2025-01-06 14:04:35,971 - INFO - MADAMIRA server stopped successfully
Lemmatizing: 100%|██████████| 1/1 [00:10<00:00, 10.97s/it]
2025-01-06 14:04:35,973 - INFO - Processing complete. Success rate: 1/1 (100.0%)
هَجْمَة صَلِيبِيّ
Topic: المنظمة الخاصة OS
Content: تصير للثورة تجميع السلاح التدريب العسكري إقناع الشعب بحتمية الكفاح المسلح قامت المنظمة بعدة عمليات فدائية اكتشفها الاستعمار في 08 03 1950
Cosine Similarity: 0.2458

Topic: المنظمة الخاصة OS
Content: تصير للثورة تجميع السلاح التدريب العسكري إقناع الشعب بحتمية الكفاح المسلح قامت المنظمة بعدة عمليات فدائية اكتشفها الاستعمار في 08 03 1950
Cosine Similarity: 0.2458

Topic: العالم السلامي في ظل الخلافة الإسلامية العثمانية:
Content: وحدة و الالتحام الأمر الذي سيجعل الدول الأوربية العاجزة عن مواجهتها العسكرية تعمل على نخرها من الداخل اعتمادا على هذه البنية لمجتمع الخلافة
Cosine Similarity: 0.1948
```

- **BM25 on Lemmatized Content:**

```
query = "الهجمات الصليبية"
search_bm25(query, data, bm25_l, lemma=True)
```

Python

```
2025-01-06 17:39:19,814 - INFO - Processing 1 texts in 1 batches
Lemmatizing:   0%|          | 0/1 [00:00<?, ?it/s]2025-01-06 17:39:19,817 - INFO - Starting MADAMIRA server...
2025-01-06 17:39:29,819 - INFO - MADAMIRA server started successfully
2025-01-06 17:39:29,821 - INFO - Input XML file generated successfully: input_0.xml
2025-01-06 17:39:30,794 - INFO - Stopping MADAMIRA server...
2025-01-06 17:39:30,879 - INFO - MADAMIRA server stopped successfully
Lemmatizing: 100%|██████████| 1/1 [00:11<00:00, 11.06s/it]
2025-01-06 17:39:30,881 - INFO - Processing complete. Success rate: 1/1 (100.0%)
Processed Query: هَجْمَة صَلِيبِيّ
Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
BM25 Score: 7.1105

Topic: علاقات الجزائر
Content: لول الجزائري الأسبان و البرتغال في السواحل المغربية و التونسية و طرابلس و كان له الفضل في هزيمة البرتغاليين في مةقعه وادي المخازن المولك
BM25 Score: 7.1105

Topic: درس حول الأسطول والبحرية الجزائريةللسنة الثالثة متوسط
Content: على اعتناق المسيحية الإتاوات مبالغ مالية تدفعها السفن البحرية الأوربية العابرة للبحر الأبيض المتوسط مقابل حماية الأسطول البحري الجزائري لها
BM25 Score: 6.8033
```

TF-IDF did not retrieve relevant documents in the lemmatized content. In this regard, it is worth mentioning that lemmatization reduced the size of the vocabulary by about 75%, which probably had great effects on the TF-IDF scores. TF-IDF is heavily dependent upon the frequency count of certain terms within the vocabulary; a huge shrinkage in the size of the vocabulary dilutes its discriminating ability to describe the differences among documents. While at the same time, BM25 yields significantly better results-the same content in the lemmatized dataset appeared, but this time, there are different rankings. BM25 appears more robust to this preprocessing step as it is a probabilistic retrieval model, incorporates term frequencies and document length normalization in a more nuanced way, which may make it less sensitive to the vocabulary reduction caused by lemmatization

- **Drawbacks of Word Count Methods:**

  Word count-based methods rely heavily on the exact words present in the text. In the case of deeper queries or richer languages, the words may vary while the context remains the same, or at least shares a similar meaning. In other words, these methods struggle to capture contextual relationships between words, even if they are relevant. To address this limitation and better access the contextual relationships between words, we will now focus on embedding model methods.

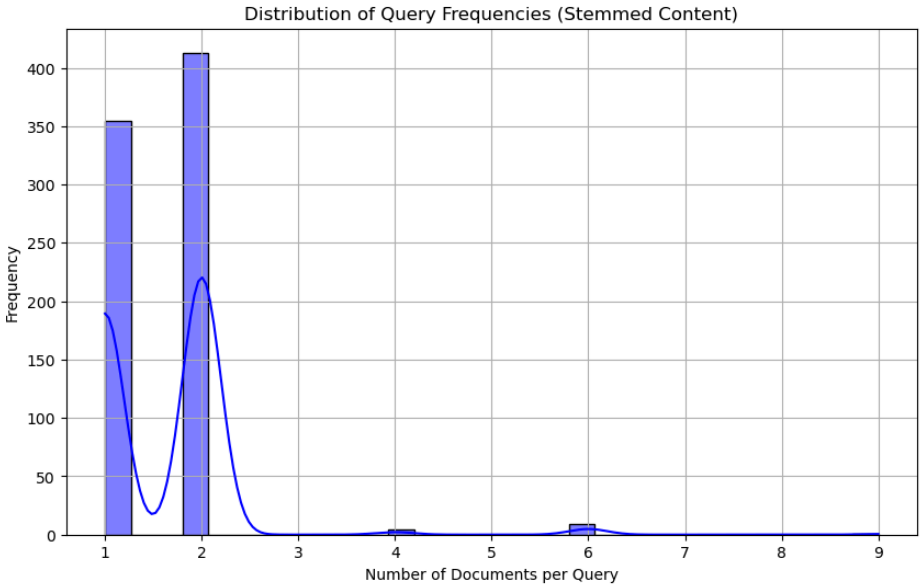


Figure.7: Distribution of Number of relevant Documents per Query

Model Embedding Methods:

Word2vec: We utilized Word2Vec to learn embeddings for all four versions of the content. A grid search was conducted to identify the optimal parameters for the Word2Vec model using the Original Content. Similar to the previous method, we extracted queries of varying lengths from the document and evaluated the performance of each model. The best model was selected based on Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) metrics

After Hypertuning the model parameters those are the best one :

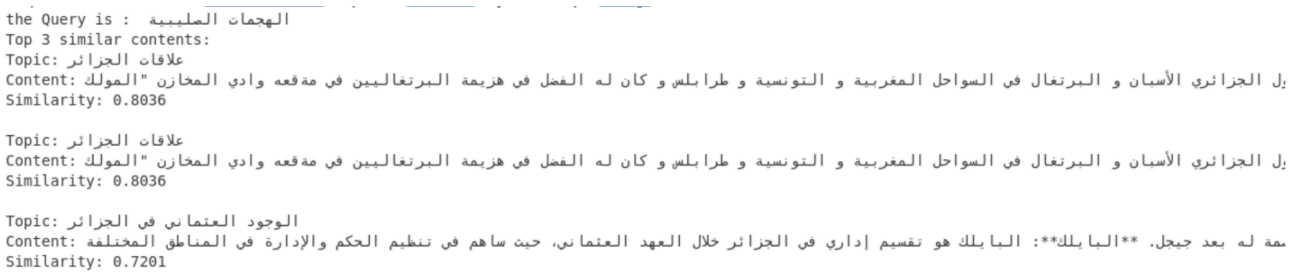
- Window Size: **200**, Min Count: **1**, Vector Size: 200, SG: **1**

And achieved those result:

- Precision: **0.4399**
- Recall: **0.8819**, MRR: **0.8779**
- Cosine Similarity Avg: **0.8591**, NDCG: **0.8603**, MAP: **0.8530**

Compared to the word count method, this approach demonstrated a significant improvement in the MAP metric, achieving around 85%, whereas the best value obtained using word count methods was 72%. This improvement suggests that the embedding model is more effective at ranking relevant documents higher, highlighting its superior ability to retrieve relevant content.

- **Try the same query:**



In this approach, the retrieved documents are all relevant to the query context. Notably, the cosine similarity scores are significantly higher compared to those obtained with TF-IDF, indicating a more accurate representation of document relevance based on semantic relationships rather than mere word frequency

- **Drawback of Word2vec(Static Embedding):**

  Word2Vec generates a unique vector for each word, irrespective of the context in which it appears. This static embedding approach becomes problematic when dealing with historical data or contexts where the meaning of terms, dates, and personalities can vary. For instance, the word "**حرب**" can have different meanings depending on the surrounding context, yet Word2Vec would assign it the same vector in all situations. This lack of context-awareness limits the model's ability to capture nuanced meanings. To overcome this limitation, we will transition to dynamic embeddings using transformer-based models, which provide context-dependent representations of words.

Transformers Models Embedding: We tested seven dynamic embedding models on the four versions of the content. The models evaluated include:

- "Omartificial-Intelligence-Space/Arabic-Triplet-Matryoshka-V2"
- "embaas/sentence-transformers-e5-large-v2"

- "embaas/sentence-transformers-multilingual-e5-base"
- "Alibaba-NLP/gte-multilingual-base"
- "intfloat/multilingual-e5-large"
- "jinaai/jina-embeddings-v3"
- "Omartificial-Intelligence-Space/GATE-AraBert-v1"

We followed the same test methodology and those are the best results acihved by each model in between the 4 version of Content:

Model	Metric	Precision	Recall	MRR	nDCG	MAP
Omartificial-Intelligence-Space/Arabic-Triplet-Matryoshka-V2	Stemmed Tashaphyne Queries	0.41	0.82	0.83	0.80	0.79
embaas/sentence-transformers-e5-large-v2	Stemmed Tashaphyne Queries	0.15	0.29	0.44	0.27	0.27
embaas/sentence-transformers-multilingual-e5-base	Stemmed Tashaphyne Queries	0.36	0.75	0.77	0.72	0.71
Alibaba-NLP/gte-multilingual-base	Stemmed Tashaphyne Queries	**0.47**	**0.94**	**0.94**	**0.93**	**0.93**
intfloat/multilingual-e5-large	Stemmed Tashaphyne Queries	0.42	0.85	0.86	0.83	0.82
jinaai/jina-embeddings-v3	Stemmed Tashaphyne Queries	0.40	0.81	0.81	0.78	0.77
Omartificial-Intelligence-Space/GATE-AraBert-v1	Stemmed Tashaphyne Queries	0.41	0.81	0.82	0.79	0.78

* Among the models tested, **`Alibaba-NLP/gte-multilingual-base`** achieves the highest performance across all metrics, making it the standout model. For **stemmed Tashaphyne queries, i**t achieves an MRR of **0.94** and MAP of **0.93**, which is the highest among all tested configurations. Other metrics for this configuration also perform exceptionally well: Precision: 0.47, Recall: 0.94, nDCG: 0.93, underscoring its effectiveness in ranking relevant results.

The second-best performance in terms of MRR is observed with the `**intfloat/multilingual-e5-large model**` using stemmed **Tashaphyne queries,** achieving an MAP of **0.82** . This configuration also shows strong performance in other metrics, including MRR: 0.86, Precision: 0.42, Recall: 0.85 and nDCG: 0.83, indicating its capability in ranking quality, though slightly behind the Alibaba-NLP model.

Results:

- BM25 significantly improved the performance of lemmatized content, which previously showed very poor results with the TF-IDF approach.
- The Word2Vec embedding model outperformed the traditional count-based word methods in terms of MAP, showcasing its superior capability in capturing semantic relationships.
- The standout model was **Alibaba-NLP/gte-multilingual-base**, which achieved the highest MAP score among all the tested methods and models, establishing itself as the most effective approach in this evaluation.

Fine Tuning Analysis:

We adopted an unsupervised fine-tuning approach, focusing on predicting the next word as the primary task. The objective was to fine-tune a larger Gemma 2 (9B) model, but due to resource limitations, we opted to fine-tune a smaller version of the model with only 2 billion parameters.

The dataset used for fine-tuning consisted of text collected from books and Databooks, which underwent a rigorous cleaning process to ensure high-quality training data

The goal of the fine-tuning process was to ensure that the model retains its general language understanding while acquiring specialized expertise in the **historical domain**. This is particularly crucial in scenarios where the context provided by RAG (Retrieval-Augmented Generation) is insufficient or entirely absent

- **Test A simple Query with Gemma 2 2b:**

```
# Test the model with a sample input
sample_text = "ما هو الدور الذي لعبه الأمير عبد القادر في مقاومة الاستعمار الفرنسي في الجزائر خلال القرن التاسع عشر؟"
inputs = tokenizer(sample_text, return_tensors="pt").to(model.device)

# Generate output
with torch.no_grad():
    outputs = model.generate(**inputs, max_length=512)

# Decode and print the output
generated_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
print("Generated Text:\n", generated_text)
```

[8] Python

```
Generated Text:
ما هو الدور الذي لعبه الأمير عبد القادر في مقاومة الاستعمار الفرنسي في الجزائر خلال القرن التاسع عشر؟

**أ - عضو في جماعة المقاومة**

**ب - رئيس تحرير صحيفة "الجزائر"**

**ج - أحد أهم قادة المقاومة**

**د - مُساعد لـ "الجزائر" في نشر الدعاية**

**الإجابة الصحيحة هي (ج) - أحد أهم قادة المقاومة**

**البيانات:**

* **الجزائر** كانت تحت سيطرة الاستعمار الفرنسي في القرن التاسع عشر.
* **أمين القادر** كان أحد أهم قادة المقاومة الجزائرية.
* **الجزائر** كانت تحارب الاستعمار الفرنسي.

**توضيح:**

* **أمين القادر** كان قيادياً في المقاومة الجزائرية، وشارك في العديد من المعارك ضد الاستعمار الفرنسي.
* **الجزائر** كانت تحت سيطرة الاستعمار الفرنسي، وشارك **أمين القادر** في مقاومة الاستعمار.
```

The generated text contains several issues: the name "أمين القادر" is incorrect and should be "الأمير عبد القادر," the explanation is overly general and repetitive, failing to provide specific details about Emir Abdelkader's significant role and achievements in resisting French colonization, and the overall response lacks depth, offering shallow insights into the historical context

We fine-tuned the Gemma 2 2B model using LoRA (Low-Rank Adaptation) with an 8-bit optimizer for efficient training. The dataset was tokenized with a maximum length of 128 tokens, and specific modules were targeted for fine-tuning. The training was conducted for 15 epochs using a batch size of 4, gradient accumulation, and a learning rate of 1e-6, with checkpoints saved at the end of each epoch. This process allowed the model to adapt to the dataset while retaining computational efficiency

Figure.8: Training Loss over 15 epochs

In Figure**.8** the training loss over 15 epochs shows fluctuations, with periods of increase and decrease. However, in the long run, there is a noticeable overall reduction in the loss across epochs.

- **Test the Query with Gemma 2 2b Fine Tuned:**

```
# Load the fine-tuned model and tokenizer
output_dir = "gemma2_2b_finetuned"  # Directory where the fine-tuned model is saved
tokenizer = AutoTokenizer.from_pretrained(output_dir)
model = AutoModelForCausalLM.from_pretrained(output_dir, device_map="auto", torch_dtype=torch.float16)

sample_text = "ما هو الدور الذي لعبه الأمير عبد القادر في مقاومة الاستعمار الفرنسي في الجزائر خلال القرن التاسع عشر؟"
inputs = tokenizer(sample_text, return_tensors="pt").to(model.device)
with torch.no_grad():
    outputs = model.generate(**inputs, max_length=512)
generated_text = tokenizer.decode(outputs[0], skip_special_tokens=True)
print("Generated Text:\n", generated_text)
```

```
Loading widget...
/usr/local/lib/python3.10/dist-packages/peft/tuners/tuners_utils.py:543: UserWarning: Model with `tie_word_embeddings=True` and the tied_targe
t_modules=['lm_head'] are part of the adapter. This can lead to complications, for example when merging the adapter or converting your model t
o formats other than safetensors. See for example https://github.com/huggingface/peft/issues/2018.
  warnings.warn(
Generated Text:
 ما هو الدور الذي لعبه الأمير عبد القادر في مقاومة الاستعمار الفرنسي في الجزائر خلال القرن التاسع عشر؟

**الإجابة:**

أمين عبد القادر كان أحد أهم القادة المقاومة الجزائرية في القرن التاسع عشر، حيث لعب دورًا هامًا في مقاومة الاستعمار الفرنسي.

**بعض أهم دور الأمير عبد القادر:**

* **قيادة المقاومة:** قاد عبد القادر العديد من المقاومات الشعبية ضد الاستعمار الفرنسي، بما في ذلك معركة عين الحص في 1832.
* **إشراك الشعب:** دعا عبد القادر الشعب الجزائري للانضمام إلى المقاومة، وشارك الشعب الجزائري في العديد من المعارك.
* **إرثه:** ترك عبد القادر إرثًا كبيرًا في مقاومة الاستعمار الفرنسي، حيث شكلت المقاومة التي قادها تأثيرًا كبيرًا على حركة التحرير الجزائرية.

**ملاحظات:**

* كان عبد القادر شخصية قوية ومثيرة للجدل، حيث كان يرى أن الجزائر يجب أن تكون دولة مستقلة.
* لعب دورًا مهمًا في تحويل المقاومة إلى حركة سياسية وسياسية.
* كان عبد القادر شخصية حكيمة وذات رؤية واضحة، حيث كان يدرك أهمية التحرير من الاستعمار.
```

Comparison and Analysis of Model Responses: Pre- and Post-Fine-Tuning

Structure and Clarity

- **Pre-Fine-Tuning:** The generated response initially reiterates the query verbatim, offering no original insight. The structure follows a multiple-choice format, unsuitable for an explanatory historical question. The correct answer, labeled as **"ج) (أحد أهم قادة المقاومة"**, is provided with minimal contextual support. Redundancies, such as repeating **"الجزائر كانت تحت سيطرة الاستعمار الفرنسي"**, detract from the response's clarity and depth. The tone resembles a quiz-style answer, misaligned with a user seeking detailed historical explanations.
- **Post-Fine-Tuning:** The refined response directly addresses the query without unnecessary repetition. It adopts a narrative tone suitable for historical analysis, organized into distinct sections: a concise answer, key contributions, and additional contextual notes. This enhances the clarity, structure, and readability of the response.

Depth of Information

- **Pre-Fine-Tuning:** The information provided is superficial, with generic statements like **"الجزائر كانت تحارب الاستعمار الفرنسي"** and lacks meaningful insights. Specific events or Emir Abdelkader's broader contributions are absent, resulting in a shallow response.
- **Post-Fine-Tuning:** The response incorporates detailed historical content, including:
- Specific examples of Emir Abdelkader's leadership, such as the **Battle of Ain El Hassin (1832)**.
- His pivotal role in mobilizing resistance efforts and unifying Algerian tribes.
- His enduring legacy in Algeria's independence movement.

  Additional insights into his political vision, cultural impact, and personal character further enrich the analysis.

Historical Accuracy

- **Pre-Fine-Tuning:** The response contains inaccuracies and irrelevant multiple-choice options, such as **"رئيس تحرير صحيفة الجزائر"** and **"مساعد لـ الجزائر في نشر الدعاية"**, which misrepresent Emir Abdelkader's role. The historical details are vague, failing to reflect the complexity of his contributions.
- **Post-Fine-Tuning:** The fine-tuned model provides historically accurate and nuanced information. Emir Abdelkader's resistance leadership, strategic acumen, and influence on Algerian history are presented with clarity and precision.

Relevance and User Satisfaction

- **Pre-Fine-Tuning:** The response fails to provide depth or specificity, includes irrelevant options, and repeats the query unnecessarily. This diminishes user satisfaction, particularly for queries requiring in-depth historical analysis.
- **Post-Fine-Tuning:** The refined response directly addresses the query, offering comprehensive insights into Emir Abdelkader's contributions and placing them within a broader historical context. The information is both relevant and tailored to user needs, significantly enhancing engagement and satisfaction.

Language Quality

- **Pre-Fine-Tuning:** While the language is formal, it lacks the sophistication expected in a historical analysis. Redundancies and repetition compromise the overall quality of the response.
- **Post-Fine-Tuning:** The language is formal, polished, and precise, maintaining an academic tone. Logical flow and the absence of repetition improve readability and engagement, ensuring the response aligns with scholarly expectations.

Conclusion

The pre-fine-tuning model generated a generic, quiz-like response with minimal historical depth, structural inadequacies, and critical inaccuracies. In contrast, the post-fine-tuning model delivers a well-structured, informative, and historically accurate response, demonstrating an improved understanding of Algerian history and Emir Abdelkader’s contributions. The fine-tuning process significantly enhances the model's ability to generate detailed, contextually rich, and user-aligned responses.

Discussion:

The results obtained in the "Dhakirate-Al-Djazair" project show the potential of the combination of state-of-the-art NLP techniques, RAG, and gamification for overcoming some of the challenges related to teaching Algerian history. The discussion now proceeds with the implications of these results, describes limitations of the system, and suggests ways to improve and possible future directions.

Results Interpretation:

1. **Performance Metrics:**
    - Embedding models differed in performance significantly. The Alibaba-NLP/gte-multilingual-base model yielded the best results throughout, both in MRR and MAP, and is therefore considered very capable of returning contextually relevant content.
    - The word count approaches like BM25 were much more successful compared to the TF-IDF method, which especially struggles to cope with the lemmatized content. The advantage of a probabilistic retrieval approach compared to purely frequency-based approaches, therefore, seems apparent.
    - Word2Vec embeddings outperformed the traditional methods in capturing semantic relationships but were themselves bound by static features.
2. **System Features:**
    - Gamified quizzes were very effective in engaging the users in an interactive, structured manner of learning. The use of cosine similarity for evaluation allowed it to be adaptable to variability in student input with a variety of vocabulary complexity.
    - The chatbot provided responses that were accurate and contextually appropriate by integrating domain-specific datasets with RAG. This demonstrated the efficiency of combining retrieval and generation in educational applications.

Limitations:

1. **Data Preprocessing:**
    - Manual annotation was highly time-consuming for preprocessing, and the complications increased with specific tagging frameworks like Tashaphyne stemming.
    - Non-Arabic texts, like French and English historical contents also, required translation APIs, adding to inaccuracies and inconsistencies.
2. **Embedding Model Challenges**:
    - Static embeddings, such as Word2Vec, are not context-sensitive; hence, they are pretty inefficient to deal with subtle historical data.
    - Transformer-based embeddings, while performing better, are computationally intensive and require huge resources for real-time deployment.
3. **Gamification and Adaptation:**
    - This may result in some generalization issues regarding the pre-defined thresholds for similarities, such as a cosine similarity of 0.84, possibly resulting in wrong quiz system

evaluations.
- More dynamic real-time feedback mechanisms could be created in addition to added question formats.

4. **Domain-Specific Challenges:**
   - History in Algeria involves several languages, contexts, and educational levels, all of which complicate complete coverage.
   -
   - Some historical topics are devoid of structured digital data, therefore making the scope of the corpus narrower.

Possible Improvements:

1. **Better Preprocessing:**
   - Automate the tagging and annotation processes using machine learning models that were trained on a small portion of data tagged manually.
   - Advanced translation and consistency-checking algorithms for handling multilingual content better.
2. **Model Improvement:**
   - Fine-tuning transformer-based models such as Alibaba-NLP/gte-multilingual-base on Algerian historical datasets to serve better for this domain.
   -
   - Hybrid embedding techniques use both static and dynamic embeddings for better capturing context.
3. **Gamification Enhancements:**
   - Adaptive difficulty levels for quizzes allow the questions to scale in difficulty based on the students' performance in real time.
   - Multimedia such as images and videos can be integrated into the quizzes to increase engagement and learning outcomes.
4. **Data Expansion:**
   - Collaborate with historians and educators in digitizing and curating more Algerian historical resources.
   - Expanding the dataset to include oral histories and cultural narratives that are so important to Algerian heritage and are usually underrepresented

Conclusion :

The "Dhakirate-Al-Djazair" project represents a significant step forward in leveraging natural language processing and retrieval-augmented generation to modernize the teaching of Algerian history. By integrating gamified quizzes, interactive lessons, and a chatbot, the platform addresses critical gaps in traditional educational methods, providing students with tailored, engaging, and accessible learning experiences.

Key findings from the project include the effectiveness of transformer-based models like Alibaba-NLP/gte-multilingual-base in delivering high-quality, contextually relevant content, and the potential of gamification to enhance engagement and retention. The structured datasets and innovative system design highlight the project's capacity to scale across multiple educational levels, fostering a deeper appreciation of Algeria's historical heritage.

Despite its accomplishments, the project's limitations underscore the need for further improvements, such as fine-tuning models for domain-specific applications, enhancing data preprocessing workflows, and expanding the dataset to include underrepresented aspects of Algerian history. Future work will focus on optimizing the platform for scalability and accessibility, ensuring it can reach diverse audiences across Algeria and beyond.

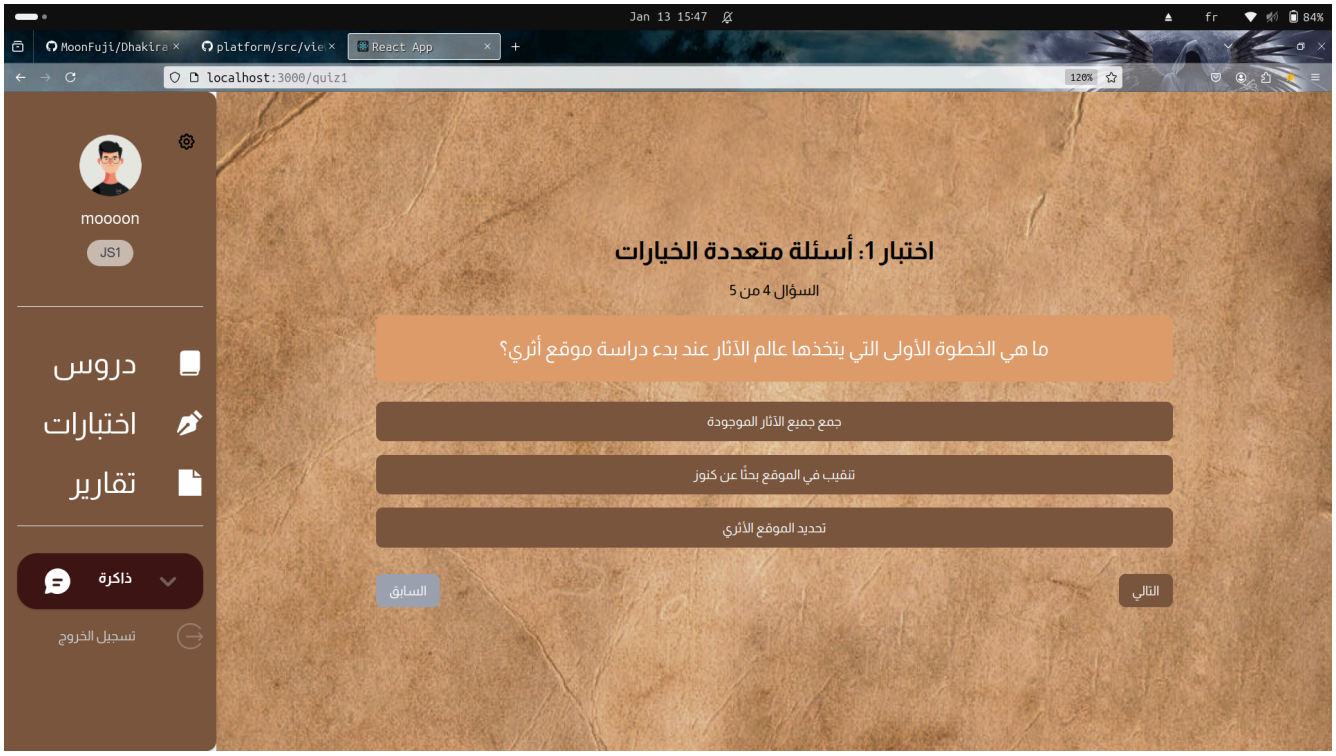
Ultimately, "Dhakirate-Al-Djazair" not only contributes to the field of educational technology but also sets the stage for broader applications of NLP-driven tools in preserving and teaching cultural heritage. By bridging historical knowledge gaps and fostering personalized learning, the project paves the way for a transformative approach to education in Algeria.

References :

Maegaard, B., Choukri, K., Cieri, C., Hamon, O., Köhler, J., Mariani, J., & Wynne, M. (2014). The language resources and evaluation conference: A forum to strengthen international collaboration. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA). https://aclanthology.org/L14-1479/

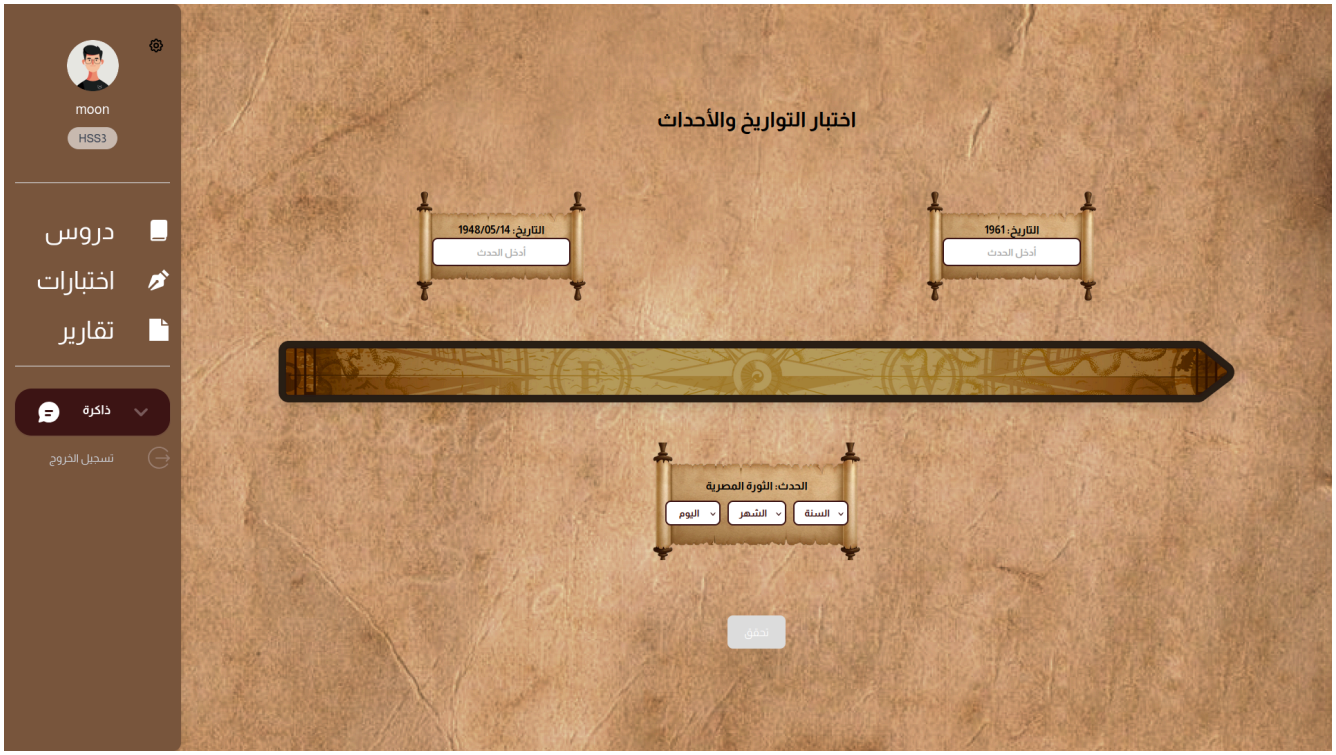
Tashaphyne. (n.d.). Tashaphyne: Arabic light stemmer library. *PyPI*. Retrieved January 13, 2025, from https://pypi.org/project/Tashaphyne/

Appendix A:



This is the quiz-1 generated by our model which represents a multiple choice question

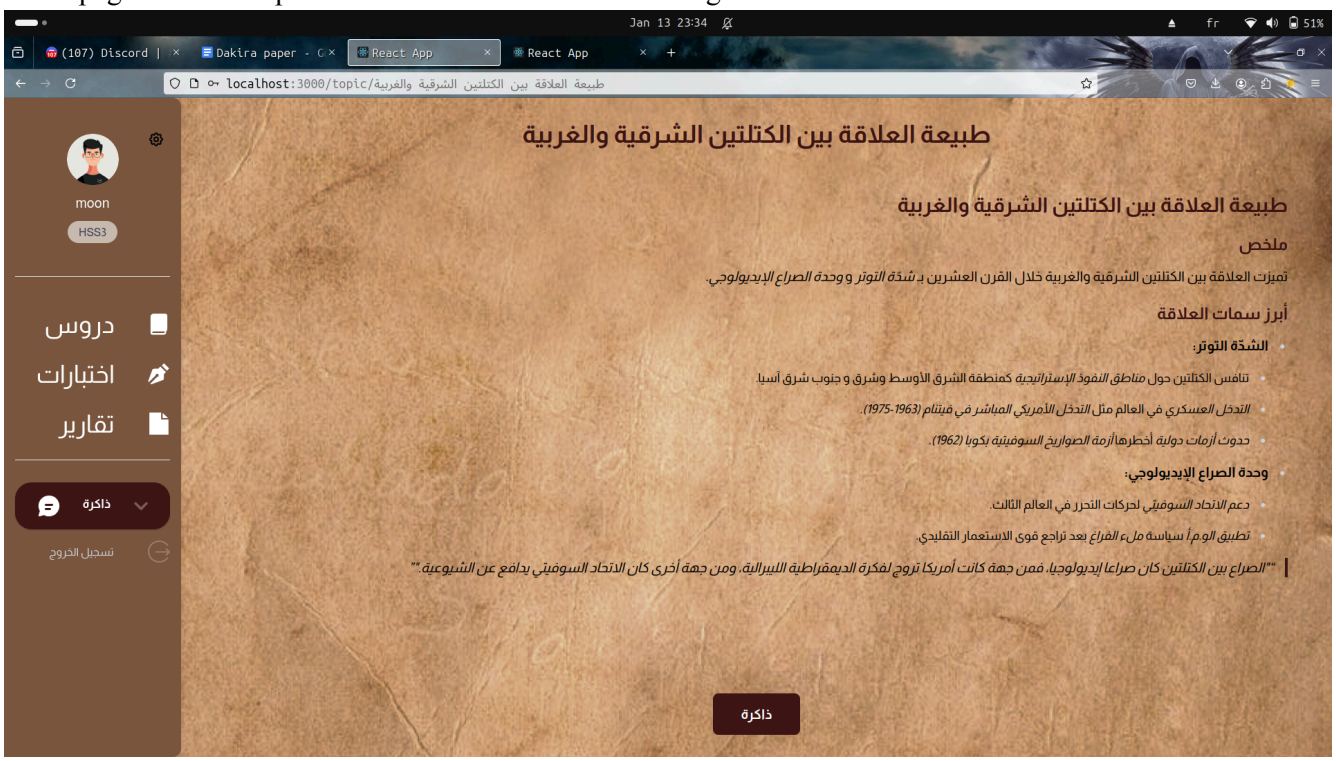
This is Quiz 3 for personality - description matching



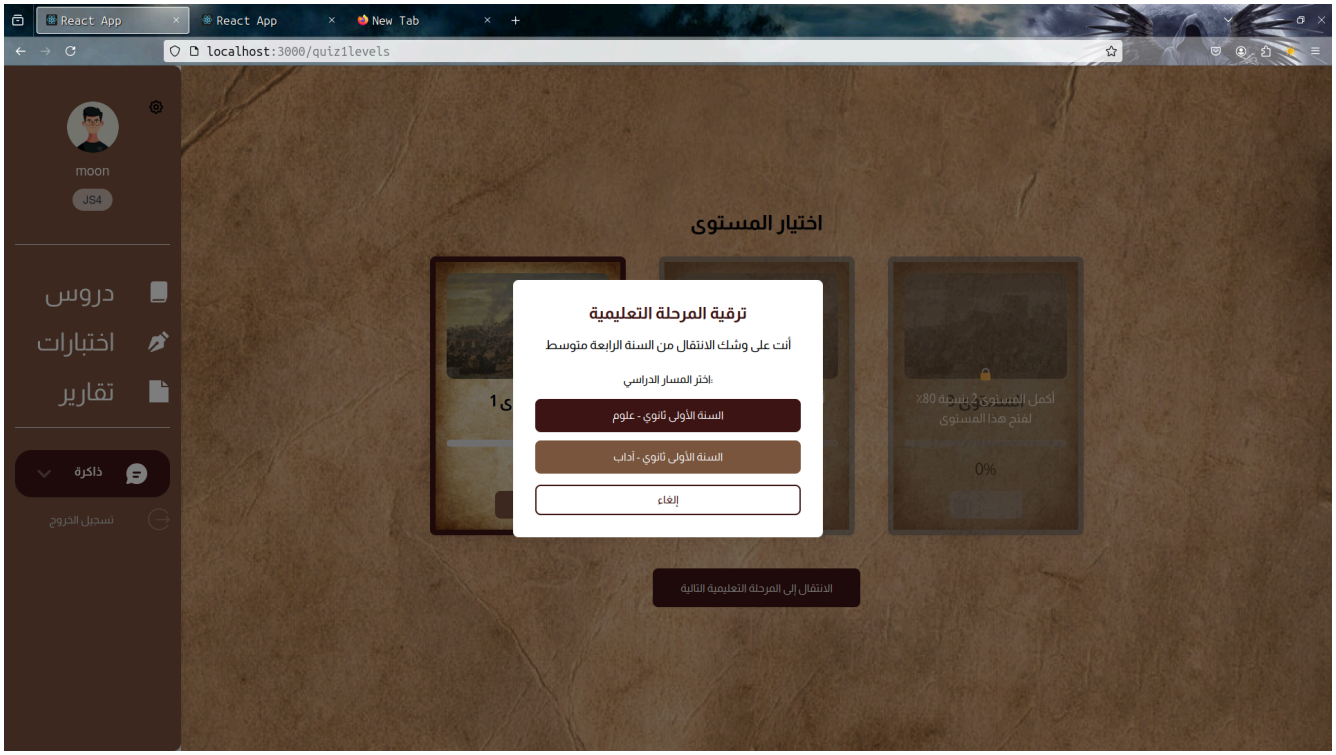
This is Quiz 2 for Date-Event matching

Chatbot page

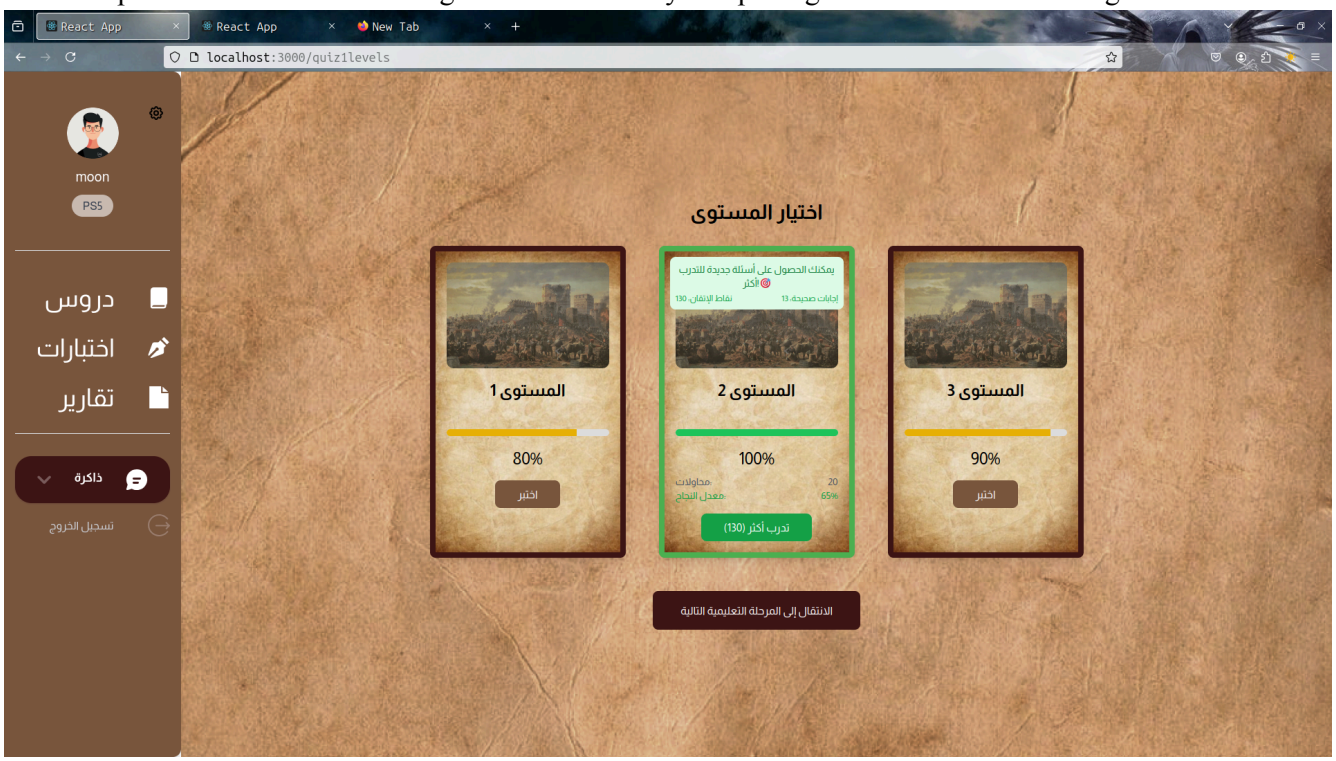
Doros page to list all topics available for an Educational Stage



Page when entering one of the ‘Doros’ which has a Lesson structured and augemented by the LLM.



Screen to pass to next educational stage after successfully completing all levels for current stage



Page for listing all levels for the current educational stage for Quiz 1 (multiple choice)