SAMSUNG

AI Course

Capstone Project
Final Report

For students (instructor review required)

Together for Tomorrow!
Enabling People
Education for Future Generations

The Intelligent Linguistic Advisor

Rannan رنّان

24/11/25

Rannan - رنّان

Abdulrazaq Al-Dawsari
Bader Alshamrani
Maher Alhijile
Raniyah Mishal Alghamdi
Rahaf Faiz Mahfoudh
Shahad Alnashwan

Content

1. Introduction

1.1. Background Information

The Arabic linguistic environment lacks reliable intelligent tools that provide accurate and fast linguistic consultations to help users understand and use the Arabic language correctly. Many people struggle to find precise linguistic answers to questions related to grammar, syntax, morphology, often resorting to complex specialized references or unreliable digital sources. As a result, linguistic errors and weak formal writing have become increasingly common in education, media, and digital content. This highlights the urgent need for a digital linguistic advisor that delivers accurate and accessible linguistic knowledge in a clear and simplified manner for all users.

1.2. Motivation and Objective

The motivation behind the Intelligent Linguistic Advisor project stems from the growing need for accessible, accurate, and fast linguistic support in Modern Standard Arabic—especially as digital communication continues to expand. Despite the richness of the Arabic language, many users struggle with grammar, morphology, spelling, and rhetorical precision, creating a clear demand for an intelligent tool that enhances writing quality and linguistic confidence.

The primary objective of this project is to develop an AI-powered conversational system capable of delivering high-quality linguistic consultations. By developing RAG Arabic language model on trusted, authoritative linguistic datasets—particularly those from the King Salman Global Academy for Arabic Language. The system aims to provide reliable linguistic guidance while reinforcing the presence of Arabic in the digital age.

1.3. Members and Role Assignments

- **Abdulrazaq Al-Dawsari (Project Manager & Data Engineer & Assistant AI Engineer)**: tracked milestones and coordinated collaboration among team members. And handled dataset collection, cleaning, and splitting,In charge of building the user interface and system integration.
- **Rahaf Mahfoudh (Data Engineer & Assistant AI Engineer)**:Focused on data preprocessing, pipeline creation, and data augmentation.
- **Raniyah Alghamdi (Data Engineer & Assistant AI Engineer)**: Focused on data preprocessing, pipeline creation, and data augmentation.
- **Maher Alhijile (AI Engineer)**: Evaluated and tested the system.
- **Bader Alshamrani (AI Engineer)**: Implemented RAG retrieval pipelines to enhance accuracy using external linguistic knowledge.
- **Shahad Alnashwan (Project Documentation Lead & Assistant AI Engineer)**: Managed timelines, created project visual identity, created WBS, created final project documents, created final project presentation, and assisted in the project overall)

Note: All team members jointly contributed to developing and optimizing the deep learning model to achieve strong performance and reliability.

1.4. Schedule and Milestones

- **Week 1**: Project initiation and planning.
- **Week 2**: Data preparation, including collection, cleaning, and preprocessing.
- **Week 3**: RAG development.

- **Week 4**: Retrieval system, backend, and frontend development, followed by testing, documentation, and final presentation. (Work is performed in parallel by team members)
- **Milestones**:
    - Action Plan drafted and submitted by end of Week 1.
    - Data collection and preprocessing completed by the end of Week 2.
    - RAG development finalized by the end of Week 3.
    - Full system implementation (Backend/Frontend) and testing completed by the end of Week 4.
    - Final project documentation and presentation delivered by the end of Week 4.

2. Project Execution

2.1. Data Acquisition

The dataset consists of Arabic grammar question answer pairs acquired from the King Salman Global Academy for Arabic Language (KSGAAL). All materials are sourced directly from the academy to ensure high quality and authenticity. Since the dataset provided by KSGAAL is focused specifically on Arabic grammar, this project works exclusively with grammar-related content.

2.2. Workflow

- **Project Initiation:** Selecting the final project idea and preparing the initial action plan.
- **Data Preparation:** Collecting grammar-focused datasets, cleaning entries, and performing preprocessing steps to ensure high-quality input.
- **RAG development:** initiating RAG development to the linguistic model to handle Arabic grammar queries effectively.
- **Retrieval System Development:** Building and integrating the retrieval component using semantic search to support accurate response generation.
- **Backend Development:** Choosing backend technologies and developing the API that connects the model with the user-facing system.
- **Frontend Development:** Creating the initial interface prototype and implementing the user interface for smooth interaction with the chatbot.
- **Testing:** Selecting user groups and conducting testing to assess accuracy, usability, and model performance across linguistic tasks.

2.3. System Design

The system was designed as a modular AI-driven architecture composed of coordinate components for data preparation, retrieval integration, model inference, and user interaction. The backend uses Python and FastAPI to manage query processing, semantic retrieval, and communication with the Qwen-based language model, enhanced through a RAG pipeline that fetches relevant Arabic linguistic knowledge before generating answers. The frontend is built using a clean and simple React interface, allowing users to submit linguistic questions and receive accurate, context-grounded responses. The entire system integrates open-

source LLMs, a vector database for retrieval, and an efficient model serving through XOONRUN, ensuring reliable and real-time Arabic linguistic consultation.

3. Results

3.1. Data Preprocessing

The dataset underwent a comprehensive pre-processing operation aimed at improving the quality and unifying the Arabic text, which is a crucial step for enhancing the efficiency of Retrieval-Augmented Generation (RAG). This process focused on applying specialized filtering procedures for the Arabic language, with an emphasis on creating two versions of the text (with and without diacritics) to support the requirements of different linguistic models. The primary tool used for text cleanup was the CAMeL Tools library.

3.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed in two stages: the first on the raw dataset, and the second on the final clean dataset, to ensure the quality of the knowledge source for the system.

3.2.1. General Overview of the Dataset and its Structure

The original dataset consisted of 8,888 question-answer pairs, which were grammatically verified by experts. The original dataset contained 9 columns: id, question, answer, domain, subdomain, instructionType, source, annotator, reviewer. Since the system relies on the Retrieval-Augmented Generation (RAG) methodology, the focus was only on the essential columns:id, question, and answer. The other columns (such as source, which was completely empty) were excluded from the embedding and indexing process.

Feature	Value	Observations
Number of Records	8,888	Question-answer pairs
Original Columns	9	id, question, answer, domain, subdomain, instructionType, source, annotator, reviewer
Used Columns (RAG)	3	id, question, answer

Unique Questions	7,788	Presence of repetition in some questions
Unique Answers	7,992	Presence of repetition in some answers

3.2.2. Data Cleaning Operations

A series of specialized cleaning operations were applied to ensure the quality and unification of the text, where each process targeted a very specific type of noise or lack of consistency in the data:

1. Remove Markdown Formatting

Strips Markdown syntax that may have been carried over from source documents (e.g., ##, **, *).

2. Remove Excel Errors

Eliminates Excel formula error strings (e.g., #NAME?, #VALUE!, #REF!).

3. Fix Punctuation Spacing

Normalizes spacing around punctuation marks (e.g., adds space after comma, removes space before question mark).

4. Replace Ampersand

Converts English ampersand & to Arabic conjunction و.

5. Remove Empty Parentheses

Deletes empty brackets () or ( ).

6. Clean Noise Markers

Removes dataset metadata accidentally included in content (e.g., "لغويات ونحو 4 ألف سؤال").

7. Clean Quranic Brackets

Removes special Quranic verse brackets﴿ ﴾ .

8. Remove Bullets and Numbering

Strips list formatting markers from line beginnings (e.g., •, -, 1., 2)).

9. Fix Punctuation Marks

Normalizes punctuation to Arabic standards (e.g., ? → ؟, , → ،).

10. Fix Spacing

Normalizes whitespace (replaces newlines with space, collapses multiple spaces, strips leading/trailing spaces).

11. Fix Spelling Errors

Corrects common Arabic spelling mistakes (e.g., إلى → الي, معنى → معني).

12. Remove Repeated Characters

Eliminates excessive character repetition (any character repeated 3+ times consecutively).

13. Clean Question-Specific

Removes question metadata (e.g., "السؤال:") (applied only to the question field).

14. Clean Answer-Specific

Removes answer metadata (e.g., "الجواب:") (applied only to the answer field).

15. Remove Diacritics

Strips all Arabic diacritical marks (applied only in the "without diacritics" track).

3.2.3.Results

During the EDA,the overall data quality and consistency improved significantly. Out of 8,888 original records, 8,294 high-quality entries were retained, reflecting an excellent data retention rate of **93.32%**. The cleaning process reduced noise, shortened overly long or redundant inputs, and improved text clarity—shown by a **98.66% text cleanliness score**. Average question lengths decreased noticeably in both characters and word count, indicating removal of unnecessary text, while answer lengths remained stable but cleaner and more consistent. Distribution plots and boxplots confirm that extreme outliers were removed, leading to tighter and more reliable data ranges. Overall, the dataset is now more structured, validated, and better aligned for developing a high-quality Arabic linguistic advisor RAG system.

```
================================================================================
📊 Comprehensive Before/After Cleaning Comparison
================================================================================

✅ Datasets loaded successfully
   📥 Before: 8888 records
   📥 After: 8294 records

================================================================================
📋 Summary Statistics Comparison
================================================================================
                    Metric Before Cleaning After Cleaning Change
             Total Records            8888           8294   -594
Avg Question Length (chars)            75.4           43.7  -31.7
  Avg Answer Length (chars)            84.7           83.6   -1.1
Avg Question Length (words)            16.1            8.3   -7.8
  Avg Answer Length (words)            15.8           15.6   -0.3
Max Question Length (chars)             766            192   -574
  Max Answer Length (chars)             760            593   -167
Min Question Length (chars)               4              4     +0
  Min Answer Length (chars)               2             10     +8
```

Figure 1 : The data size before and after cleaning

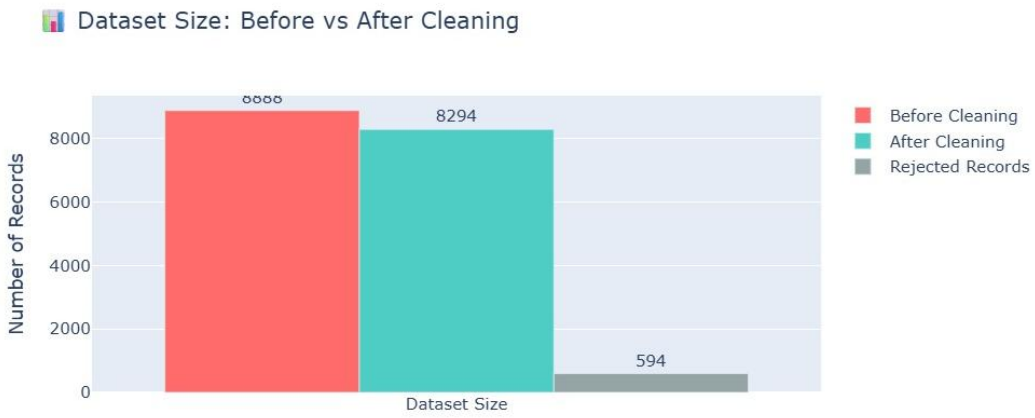

Figure 2 : The data size before and after cleaning visualization

Figure 3 : The average question length before and after cleaning visualization

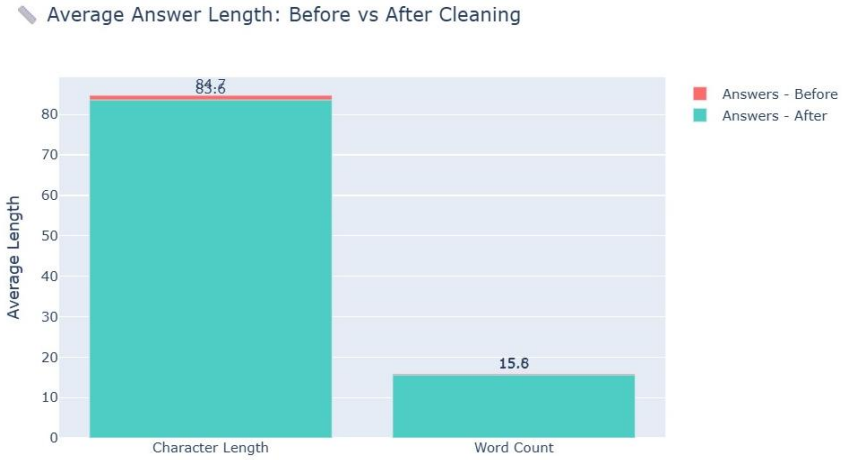

Figure 4 : The average answer length before and after cleaning visualization

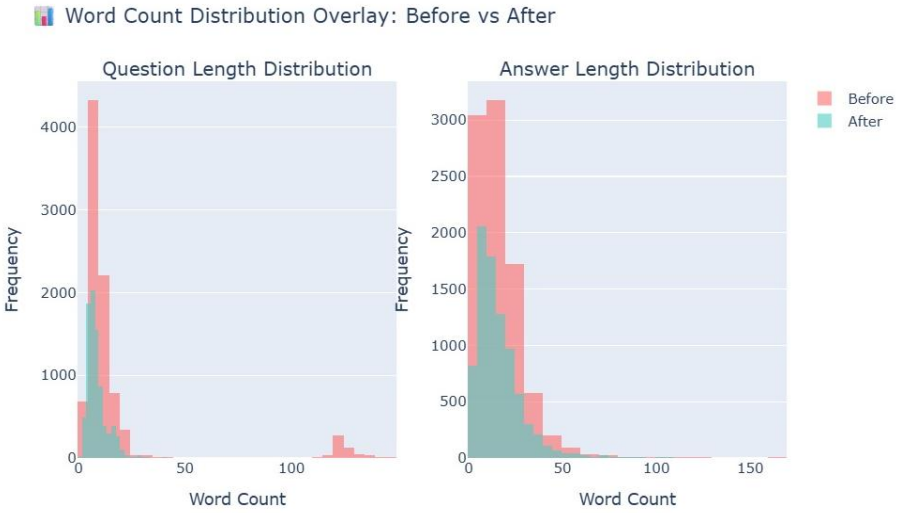

Figure 5 : The word count distribution overlay before and after cleaning visualization

Figure 6 : The quality metrics radar before and after cleaning visualization

Figure 7 : The text length distribution before and after cleaning visualization

```
================================================================================
📈 Data Quality Improvement Metrics
================================================================================
                Metric Percentage      Status
    Data Retention Rate     93.32%  ✅ Excellent
Average Text Cleanliness    98.66%  ✅ Improved
  Record Validation Rate    93.32%  ✅ High Quality
```

Figure 8 : The data quality improvements metrics

```
📍 BEFORE CLEANING:
  Q: ما نوع الإعلال في الفعل ( نأت ) وما وزنه ؟
  A: نأت: فيه إعلال بالحذف لمناسبة الجزم وأصله نأتي، وزنه نفع.

📍 AFTER CLEANING:
  Q: ما نوع الإعلال في الفعل (نأت)؟ وما وزنه؟
  A: نأت:فيه إعلال بالحذف لمناسبة الجزم وأصله نأتي، وزنه نفع.

📊 Character Count Change:
  Question: 44 → 40 (-4)
  Answer: 84 → 56 (-28)
```

Figure 9 : Before/After cleaning text examples

3.3. Modeling

The Intelligent Linguistic Advisor system was developed using a Retrieval-Augmented Generation (RAG) architecture designed to produce accurate, context-grounded answers to Arabic linguistic

questions. The modeling process began with preparing the dataset provided by the King Salman Global Academy for Arabic Language (KSGAAL), which consisted of 8,294 (after cleaning) high-quality Arabic grammar question–answer pairs stored in JSON format. Each JSON record contained an ID and both diacritized and undiacritized versions of the question and answer. To make the data suitable for retrieval, every entry was transformed into a unified text document, structured as: "Question: …\nAnswer: …". These documents formed the basis for embedding and indexing.

For semantic retrieval, embeddings were generated using Qwen3 Embedding 8B, which supports vector sizes ranging from 32 to 4096 dimensions. Each processed document was converted into a fixed-size embedding that captures its semantic meaning for similarity search. The generated embeddings were then stored in ChromaDB, a fast, locally running vector database used to support nearest-neighbor search. Embeddings were produced in batches: the application sent groups of text documents to the embedding model, received their vectors, and upserted the documents, IDs, and embeddings into a dedicated ChromaDB collection. This completed the indexing phase, enabling the system to efficiently retrieve the most relevant linguistic references for any future user query.

For answer generation, the system used Qwen 32B, a large 32-billion-parameter language model capable of strong reasoning and native Arabic understanding. The model was based on a carefully designed system prompt to ensure consistent behavior specifically instructing it to answer strictly based on retrieved context and to respond in Arabic only. During the retrieval and generation phase, the user's question is first embedded using the same embedding model, then matched against stored vectors in ChromaDB to retrieve the top relevant documents. These retrieved items are compiled into a concise context package. Finally, the system sends the context + user question to Qwen 32B, which generates a linguistically accurate answer grounded exclusively in the supplied references.

Overall, the completed RAG architecture follows a clear and reproducible workflow:
Indexing: Read JSON data → build documents → generate embeddings → store in ChromaDB.
Retrieval & Generation: User question → embed → similarity search → retrieve best matches → construct context → generate final answer with Qwen 32B.

This modeling approach ensured that the final system combined the strengths of semantic retrieval with advanced Arabic language generation, resulting in reliable, reference-backed linguistic consultation.

3.4. User Interface

A simple React web application was developed as the user interface. It allowed users to navigate easily in the website and to ask their questions related to the Arabic language.

Figure 10 : Home Page

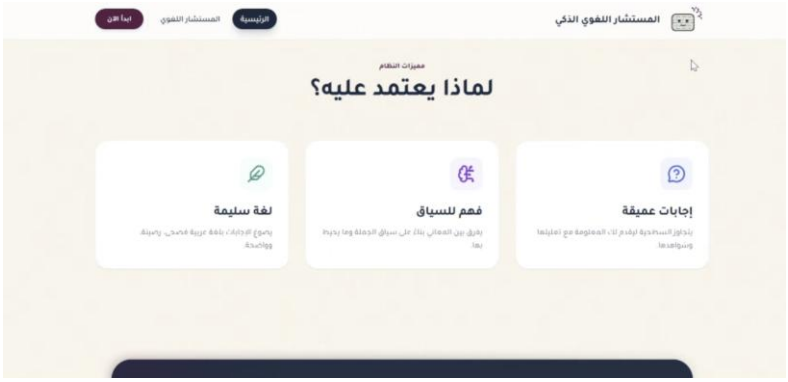

Figure 11 : Home Page, "Why can you rely on the intelligent linguistic advisor ?"

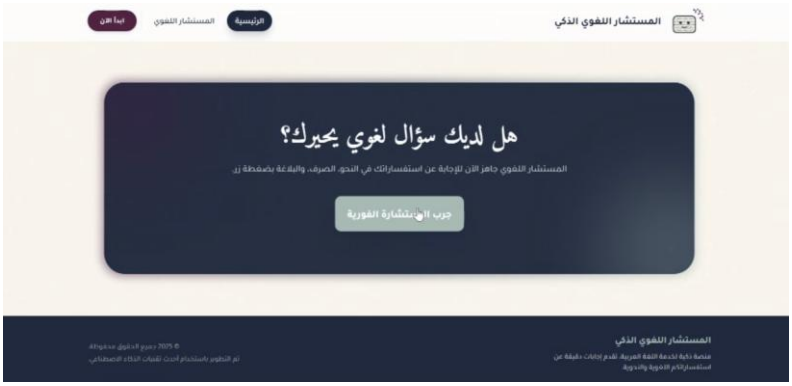

Figure 12 : Home Page, "Do you have a linguistic question ? Try a linguistic consultation"

Figure 13 : The chat bot page

3.5. Testing and Improvements

The Intelligent Linguistic Advisor underwent extensive testing using a dedicated evaluation dataset designed to measure both retrieval quality and answer accuracy. The system achieved strong overall performance, consistently understanding user queries and providing accurate, source-supported answers. During testing, several iterative improvements were applied, including enhancements to data cleaning, adjustments to preprocessing, and refinements to retrieval scoring, which reduced noise and improved the precision of retrieved contexts. These improvements contributed to the system's final high evaluation scores and a more reliable user experience. Future testing will extend to the full web interface to validate performance during real-time user interactions.

The metrics used belonged to two distinct categories, aiming to measure generated answer quality and retrieved context quality. Each metric was computed automatically using an evaluation pipeline that compared the model's outputs and retrieved texts against reference answers and annotated ground-truth labels.

To evaluate retrieval quality, the system used the model (llama-3.1-8b-instant) as the scoring engine responsible for interpreting the retrieved text and comparing it to the user's query and the gold reference answer. The model played a central role in assigning relevance judgments by analyzing semantic relationships, detecting alignment, and identifying missing or noisy information. Its strong multilingual understanding allowed it to reliably evaluate Arabic queries and retrieved contexts with minimal prompting.

In the **generated answer quality** phase the following metrics were measured:

- **Context Relevance**:
  evaluates how well the system identifies and uses the parts of the retrieved context that are genuinely pertinent to the user's query.
  Score achieved: 99%, indicating extremely strong alignment between retrieved context and query intent.

- **Answer Relevance:**
  This measures whether the final generated answer directly addresses the user's question rather than drifting or adding unnecessary information.
  Score achieved: 98%, showing that responses remained highly focused on user needs.

- **Correctness:**
  This assesses factual accuracy and whether claims in the answer are properly grounded in the retrieved sources.
  Score achieved: %92, reflecting a high level of accuracy across evaluated answers.

- **Lexical overlap**
  Indicates how much of the generated answer directly reuses wording from the retrieved text versus paraphrasing.
  Score achieved: 72%, moderate overlap, suggests healthy paraphrasing rather than excessive copying, while still staying grounded.

In the **retrieval quality phase** the following metrics were measured:

- **Recall@3**

  This metric checks whether the correct supporting information appears within the top three retrieved documents.

  Score achieved: 94%, demonstrating strong consistency in retrieving the right information early.

- **NDCG@3 (Normalized Discounted Cumulative Gain)**

  This evaluates not only whether relevant documents were retrieved, but also whether they were ranked highly among the top three results.

  Score achieved: %94, indicating that the system not only found the right documents but ordered them effectively.

- **Context precision**

  Evaluates how clean and focused the retrieved context is, reflecting how much irrelevant or noisy information remains.

  Score achieved: 79%, generally good, but reflect that some extra or unnecessary context still appears, leaving room for refinement.

Examples:

Below shown multiple random samples taken from evaluation results:

Sample 282

Question:

ما الحكم إن كان "فعاليل" ثلاثيًا مزيدًا بحرفين؟

Reference:

ففي هذه الحالة يُحذف حرف واحد، نحو: منطلق → مطالق.

Model Answer:

ففي هذه الحالة يحذف حرف واحد، نحو: منطلق: مطالق.

Retrieved Contexts (summary):

حكم فعاليل إذا كان ثلاثيًا مزيدًا بحرفين: يُحذف حرف واحد.
حكم فعاليل إذا كان مزيدًا بثلاثة أحرف: يُحذف حرفان.
شبه فعالل قياسي في كل اسم ثلاثي مزيد.

Metrics:

Context: relevance 3, precision 2
Answer: relevance 3, faithfulness 3, correctness 3, lexical_overlap 1.0
Retrieval: recall@3 1.0, precision@3 0.265, MRR 1.0, NDCG@3 1.0
Final score: 2.8

Comment:

الجواب صحيح ولكن لم يحتوي على تفاصيل إضافية.

Sample 106

Question:

ما يقال عند ذكر أدوات يكثر دورها في الكلام؟

Reference:

يقال في (لم): حرف جزم لنفي المضارع وقلبه ماضيًا، ويزاد في (لما) النافية: "متصل نفيه، متوقع ثبوته".

Model Answer:

عذراً، لا توجد معلومات كافية في البيانات للإجابة على هذا السؤال.

Retrieved Contexts (summary):

ثم: حرف عطف للترتيب والمهلة
قد: حرف تحقيق وتوقع وتقليل
"كلا: حرف ردع وزجر أو بمعنى "حقًا

Metrics:

Context: relevance 3, precision 2
Answer: relevance 3, faithfulness 0, correctness 0, lexical_overlap 0.037
Retrieval: recall@3 0.176, precision@3 0.12, MRR 0.0, NDCG@3 0.0
Final score: 1.6

Comment:

النص المنسوب غير موجود في البيانات المسترجعة، ولذلك الإجابة غير مدعومة.

Sample 275

Question:

أين يكون المجاز المرسل، والاستعارة؟

Reference:

المجاز المرسل يكون في الكلمة، والاستعارة تكون في الكلام.

Model Answer:

المجاز المرسل يكون في الكلمة، والاستعارة تكون في الكلام.

Retrieved Contexts (summary):

نفس الجواب الوارد في المرجع
الفرق بين المجاز المرسل والاستعارة
مثال على علاقة السببية في المجاز المرسل

Metrics:

Context: relevance 3, precision 3
Answer: relevance 3, faithfulness 3, correctness 3, lexical_overlap 1.0
Retrieval: recall@3 1.0, precision@3 0.163, MRR 1.0, NDCG@3 0.973
Final score: 3.0

Comment:

الجواب صحيح ومتوافق تمامًا مع النصوص
.

```
============================================================
EVALUATION METRICS SUMMARY (Scale 0-1)
============================================================

Metric                          Mean            Std
------------------------------------------------------------
context_relevance               0.990           0.090
context_precision               0.787           0.181
answer_relevance                0.980           0.132
faithfulness                    0.925           0.220
correctness                     0.923           0.224
lexical_overlap                 0.723           0.327

============================================================
RETRIEVAL METRICS SUMMARY (Scale 0-1)
============================================================
```

Figure 14 : The evaluation metrics summary

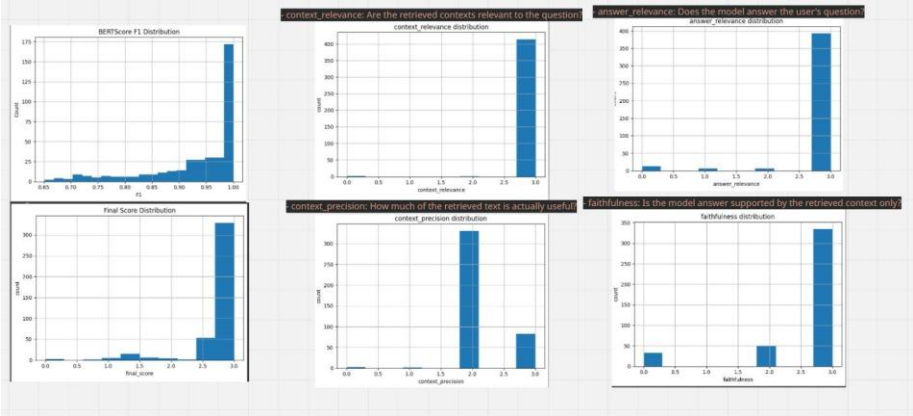

Figure 15 : The evaluation metrics visualization

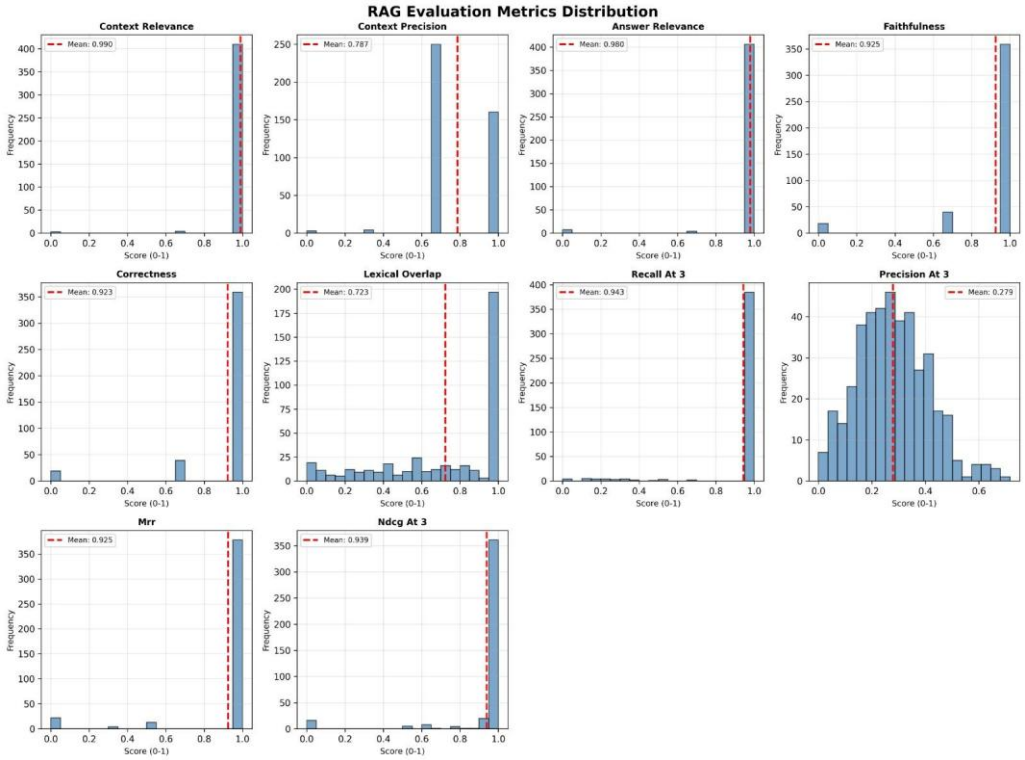

Figure 16 : The RAG evaluation metrics distrubution visualization

4. Projected Impact

4.1. Accomplishments and Benefits

The Intelligent Linguistic Advisor project achieved several key milestones that significantly advanced the development of an accessible and academically reliable Arabic language assistant. One major accomplishment was the successful creation of a complete RAG-based system capable of delivering real-time, reference-grounded linguistic guidance using authoritative data from KSGAAL. The project also produced a fully functional web interface that allows users to interact with the advisor seamlessly, demonstrating the practical value of integrating retrieval systems with modern user-facing applications. In addition, the team implemented a comprehensive data-cleaning and preprocessing pipeline that improved dataset quality and ensured consistent linguistic structure across thousands of entries. Beyond meeting the initial objectives, the project

highlights the feasibility of applying retrieval-augmented methods to support Arabic grammar, morphology, and spelling queries—offering a meaningful contribution to the development of digital linguistic tools for researchers, students, and Arabic language enthusiasts.

4.2. Future Improvements

The team members wish to continue improving the model in collaboration with KSGAAL, utilizing fine-tuning techniques to expand the model's scope and enhance its responses for greater accuracy. The focus will be on further customizing the model to cover a broader range of linguistic topics, while enhancing its responses to better align with context, thus improving the model's effectiveness in handling grammar, morphology, and spelling queries.

5. Appendix

- Schedule Summary :
  - To open WBS Click Here
  - The linguistic advisor project Click Here
- Website walkthrough video Click Here
- The Linguistic Advisor project on GitHub Click Here
- The Website Link Click Here

6. Team Member Review and Comment

NAME	REVIEW and COMMENT
Rahaf Faiz mahfoudh	I am grateful to have been part of this rewarding experience and enjoyed working with a dedicated team that made it even better.
Raniyah Alghamdi	It was a great experience working on this project with such a supportive team.
Shahad Alnashwan	It was a great honor learning and working with my colleagues.
Abdulrazaq Al-Dawsari	Let's stay in touch and build our network on LinkedIn
Bader Alshamrani	This experience meant a lot to me, and it's all thanks to a team that made me feel supported, challenged, and genuinely connected every step of the way.
Maher Alhijile	Great experience with a lot of growth. Honored to have worked on this project alongside my colleagues.

7. Instructor Review and Comment

CATEGORY	SCORE	REVIEW and COMMENT
IDEA	__/10	
APPLICATION	__/30	
RESULT	__/30	
PROJECT MANAGEMENT	__/10	

PRESENTATION & REPORT	__/20	
TOTAL	__/100	