ARABIC QUESTION ANSWERING BASED ON LLM

Arabic language, Question Answering, Natural Language Processing, AraElectra, Fine-Tuning, DuckDuckGo, NLP

DEVELOPMENT OF THE SYSTEM

This presentation outlines the development of an advanced Arabic Question Answering (QA) system designed to address the scarcity of robust QA tools for Arabic.

By fine-tuning the AraElectra model and integrating NLP techniques like Named Entity Recognition (NER), query reformulation, and web based retrieval via the DuckDuckGo API.

the system delivers contextually accurate answers. Evaluated on the AAQAD dataset, it achieves an F1 score of 0.4814 and accuracy of 0.6372, showing strong potential for applications in education, research, and customer service

WHY ARABIC QA MATTERS

CONTEXT

Arabic QA systems lag behind English due to linguistic complexities (e.g., morphology, diacritics) and limited annotated data

GOAL

Develop a smart QA system to extract precise answers from web documents or user inputs, enhancing information access for Arabic speakers.

APPLICATIONS

Education, research, customer service, leveraging growing Arabic online content.

KEY FEATURES

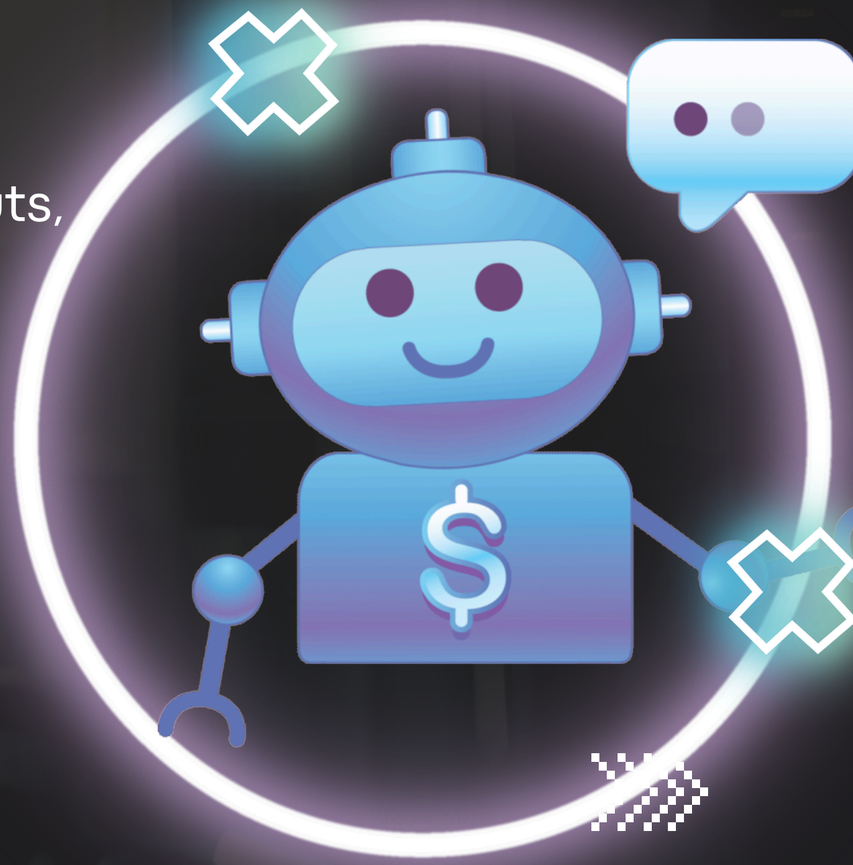
Supports bilingual input (Arabic/English), multiple context modes (predefined, manual, web-based), and a user-friendly Gradio interface.

HOW THE SYSTEM WORKS

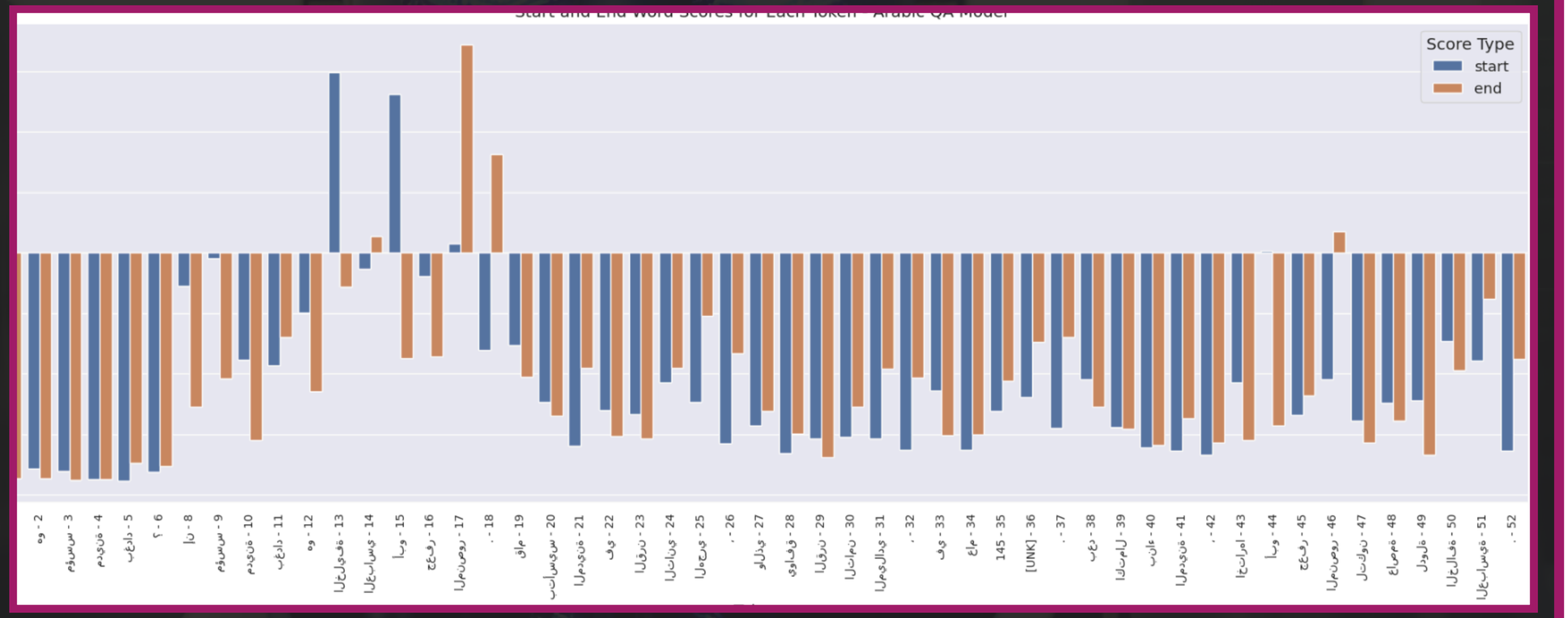
Query Processing: Preprocesses questions using tokenization, NER, and query reformulation to confirm user intent.

Context Retrieval: Supports predefined texts, user-provided inputs, or real-time web searches via DuckDuckGo API, ranked by relevance.

AraElectra : A transformer-based model optimized for Arabic, using discriminative pre-training for efficiency.



TOKENIZER AND VISUALIZING SCORES




```
306          من (Question)
583          هو (Question)
10611        مؤسس (Question)
1171         مدينة (Question)
3112         بغداد (Question)
105             ؟ (Question)

[SEP]           3 (Question)

476             إن (Context)
10611        مؤسس (Context)
1171         مدينة (Context)
3112         بغداد (Context)
583          هو (Context)
12280        الخليفة (Context)
29743        العباسي (Context)
1195         أبو (Context)
9501         جعفر (Context)
17647        المنصور (Context)
.               20 (Context)
1178         قام (Context)
22077        بتأسيس (Context)
1665         المدينة (Context)
305          في (Context)
2890         القرن (Context)
1161         الثاني (Context)
27231        الهجري (Context)
،               103 (Context)
1619         والذي (Context)
12423        يوافق (Context)
2890         القرن (Context)
6687         الثامن (Context)
22923        الميلادي (Context)
،               103 (Context)
305          في (Context)
515          عام (Context)
145          23045 (Context)
[UNK]           1 (Context)
.               20 (Context)
446          بعد (Context)
14166        اكتمال (Context)
1745         بناء (Context)
```

HOW THE SYSTEM WORKS CONT..

Answer Generation: Uses fine-tuned AraElectra model with prompt engineering (four strategies: direct, formal specificity, detailed precision, analytical completeness) for accurate answer spans.

User Interface: Gradio-based, bilingual, interactive platform for input and answer display.

BUILDING THE ARABIC QA SYSTEM

Query Processing: Handles Arabic linguistic challenges (diacritics, morphology) via AraElectra's tokenizer and NER for entity-focused answers.

Context Retrieval: Uses DuckDuckGo API for web searches, ranking documents by keyword overlap and semantic similarity.

Answer Generation: Fine-tuned AraElectra model predicts answer spans, enhanced by prompt engineering and rule-based post-processing for grammar and cultural fit.

IMPLEMENTATION

Tools: Python, Hugging Face Transformers, Gradio, DuckDuckGo API.

Dataset: AAQAD, split 80% training, 10% validation, 10% test.

Fine-Tuning: 5 epochs, AdamW optimizer, 2e-5 learning rate, cross-entropy loss.

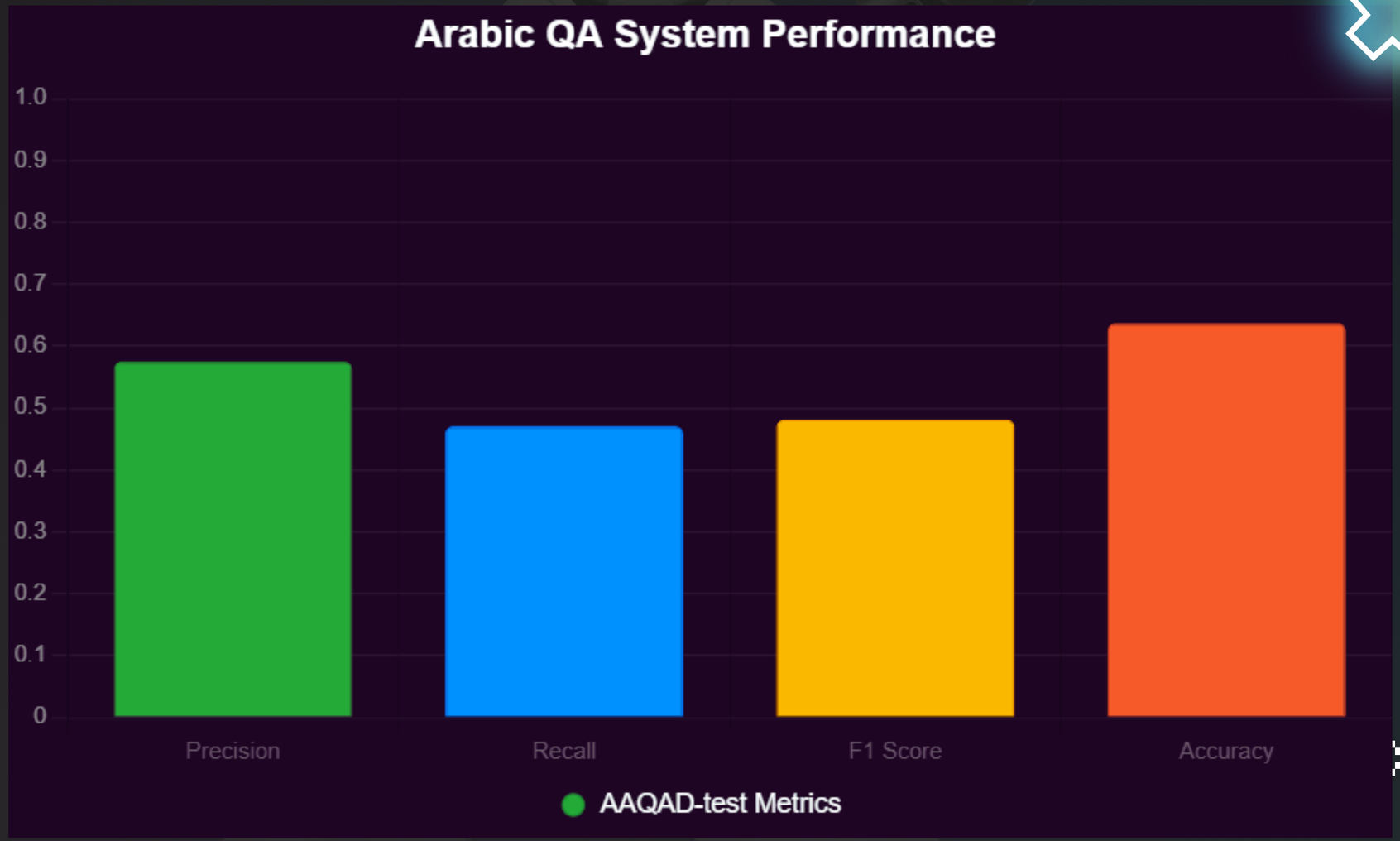
Data Augmentation: Synonym replacement, back-translation for robustness.

PERFORMANCE AND INSIGHTS

Dataset: AAQAD-test, Arabic question-answer pairs.

METRICS

PRECISION: 0.5754
RECALL: 0.4708
F1 SCORE: 0.4814
ACCURACY: 0.6372



SAMPLE PERFORMANCE

Factoid question (e.g., “What is the capital of Iraq?”): Perfect scores (F1=1.0)

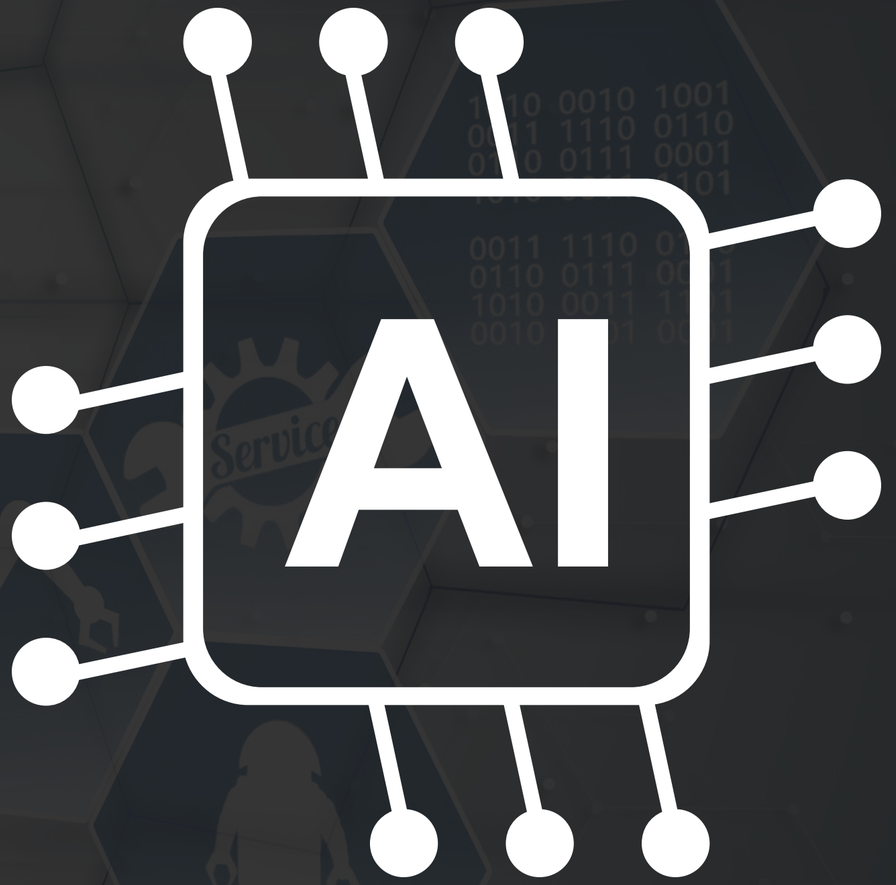
Complex questions (e.g., space exploration): Mixed results, e.g., missed “Blue Origin” for reusable rockets, vague Artemis program answers.

TESTING

Unit Testing: Validated tokenizer, answer generation

Integration Testing: Ensured seamless query-to-answer pipeline.

User Acceptance: Confirmed usability via Gradio interface, strong for factoid queries, weaker for inferential ones.



EXAMPLES

```
Technology Context:
في القرن العشرين، تطورت تكنولوجيا المعلومات بشكل مذهل. تم تطوير الكمبيوترات والإنترنت وأجهزة الهواتف
المحمولة وغيرها من الأجهزة الإلكترونية التي أصبحت جزءًا أساسيًا من حياتنا اليومية. بفضل هذه
التكنولوجيا، أصبحت الاتصالات أسرع وأسهل، وزادت سرعة نقل المعلومات والبيانات. كما تطورت البرامج
والتطبيقات التي توفر العديد من الخدمات والوظائف المفيدة للأفراد والشركات.

--------------------------------------------------
Question: ما هي التكنولوجيا التي ظهرت في القرن العشرين؟
Context length: 361 characters
--------------------------------------------------
Query has 72 tokens.

Segment A (Question) tokens: 11
Segment B (Context) tokens: 61
Answer start position: 16
Answer end position: 17
Answer tokens: ['تكنولوجيا' ,'المعلومات']
Answer: "تكنولوجيا المعلومات"
Confidence: 0.8793
--------------------------------------------------
Question: ما الأجهزة الإلكترونية التي أصبحت جزءًا من حياتنا اليومية؟
Context length: 361 characters
--------------------------------------------------
Query has 73 tokens.

Segment A (Question) tokens: 12
Segment B (Context) tokens: 61
Answer start position: 24
Answer end position: 29
Answer tokens: ['الكمبيوتر' ,'##ات' ,'والإنترنت' ,'وأجهزة' ,'الهواتف' ,'المحمولة']
Answer: "الكمبيوترات والإنترنت وأجهزة الهواتف المحمولة"
Confidence: 0.8472
--------------------------------------------------
Question: ما الفوائد التي يمكن أن تحققها هذه البرامج والتطبيقات للأفراد والشركات؟
Context length: 361 characters
--------------------------------------------------
Query has 75 tokens.

Segment A (Question) tokens: 14
Segment B (Context) tokens: 61
Answer start position: 68
Answer end position: 69
Answer tokens: ['والوظائف' ,'الخدمات']
```

Technology context

```
Context Text (wrapped):
إن عاصمة العراق هي مدينة بغداد ، كانت تسمى قديماً مدينة السلام. تقع مدينة بغداد
في وسط العراق، وهي تابعة لمحافظة بغداد. كانت بغداد قديمًا العاصمة المركزية
للخلافة العباسية، فالذي بنى مدينة بغداد هو الخليفة العباسي أبو جعفر المنصور في
القرن الثاني الهجري، والذي يوافق القرن الثامن الميلادي، وبعد اكتمال بناء
المدينة، اتخذها أبو جعفر المنصور عاصمة لدولة الخلافة العباسية، وبقيت بغداد عاصمة
الخلافة العباسية حتى جاء الخليفة العباسي المعتصم بالله، وقام ببناء مدينة سامراء،
وجعلها عاصمة الخلافة، ومدينة سامراء تقع شمال مدينة بغداد.

==================================================
Testing answer_question function:

1. First Question:
Question: ما هي عاصمة العراق؟
Context length: 526 characters
--------------------------------------------------
Query has 105 tokens.

Segment A (Question) tokens: 7
Segment B (Context) tokens: 98
Answer start position: 12
Answer end position: 12
Answer tokens: ['بغداد']
Answer: "بغداد"
Confidence: 0.8378
--------------------------------------------------

2. Second Question:
Question: ماذا كانت تسمى قديماً؟
Context length: 526 characters
--------------------------------------------------
Query has 105 tokens.

Segment A (Question) tokens: 7
Segment B (Context) tokens: 98
Answer start position: 17
Answer end position: 18
Answer tokens: ['السلام' ,'مدينة']
Answer: "مدينة السلام"
Confidence: 0.9730
--------------------------------------------------
```

Baghdad context

TESTING

```
--- Question 1 ---
Question: في أي عام تم تأسيس مدينة بغداد؟
Context length: 237 characters
--------------------------------------------------
💉 Using Advanced Prompt Engineering...

📊 Generated 4 different prompt strategies:
  • direct_None_clean: 145 (confidence: 0.6536)
  • formal_specificity_clean: 145 (confidence: 0.8449)
  • detailed_precision_structure: 145 (confidence: 0.7886)
  • analytical_completeness_segment: القرن الثاني الهجري (confidence: 0.1542)

🎯 Best Answer Selected:
Answer: 145
Confidence: 0.8449
Strategy Used: formal_specificity_clean

--- Question 2 ---
Question: في أي قرن تم بناء مدينة بغداد؟
Context length: 237 characters
--------------------------------------------------
💉 Using Advanced Prompt Engineering...

📊 Generated 4 different prompt strategies:
  • direct_None_clean: الثاني الهجري (confidence: 0.2651)
  • formal_specificity_clean: الثاني الهجري (confidence: 0.2573)
  • detailed_precision_structure: الثاني الهجري (confidence: 0.3797)
  • analytical_completeness_segment: الثاني (confidence: 0.2906)

🎯 Best Answer Selected:
Answer: الثاني الهجري
Confidence: 0.3797
Strategy Used: detailed_precision_structure
```

Prompt engineering strategies for Arabic QA

INTERFACE

Gradio-based user interface for the Arabic QA system

CURRENT CHALLENGES

Inferential Questions: Lower F1 score (0.4398) due to limited semantic reasoning

Long Contexts: Performance degrades due to memory constraints.

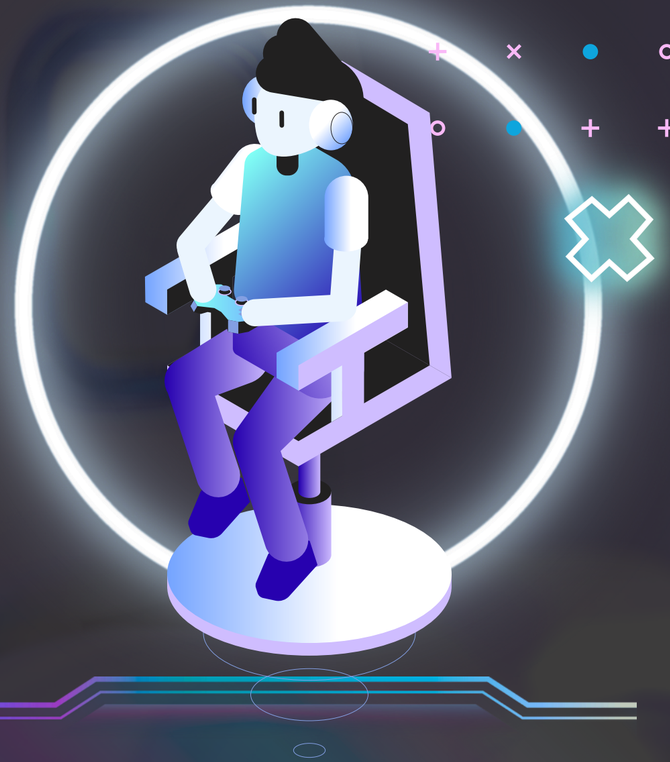
Complex Queries: Prompt engineering less effective for non-factoid questions.

Example: Space exploration questions showed incomplete or vague answers (e.g., missed “Perseverance” for Mars rovers).

ENHANCING THE SYSTEM

Long-Context Optimization: Use context chunking or sparse attention (AraBERT) to handle longer texts.

Inferential Questions: Integrate knowledge graphs or external knowledge bases for better reasoning.



Dynamic Prompts: Apply reinforcement learning for adaptive prompt generation.

Multilingual Support: Enable cross-lingual QA (Arabic-English).

User Interface: Add real-time feedback and voice query support for broader accessibility.

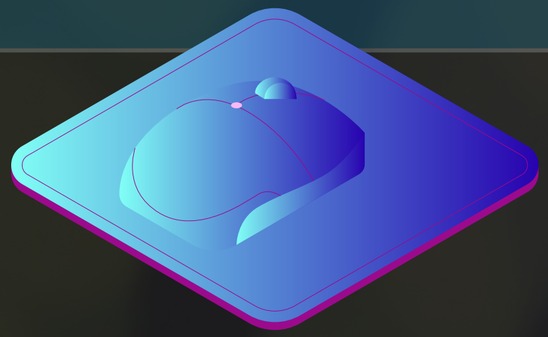
KEY TAKEAWAYS

Achievements :Built a robust Arabic QA system with AraElectra, achieving 0.4814 F1 and 0.6372 accuracy on AAQAD.

Innovations :Integrated prompt engineering, web retrieval, and Gradio interface for usability.

Impact :Enhances Arabic NLP, supporting education, research, and customer service.

Future Potential :Scalable for multilingual and complex query applications.



THANK YOU!