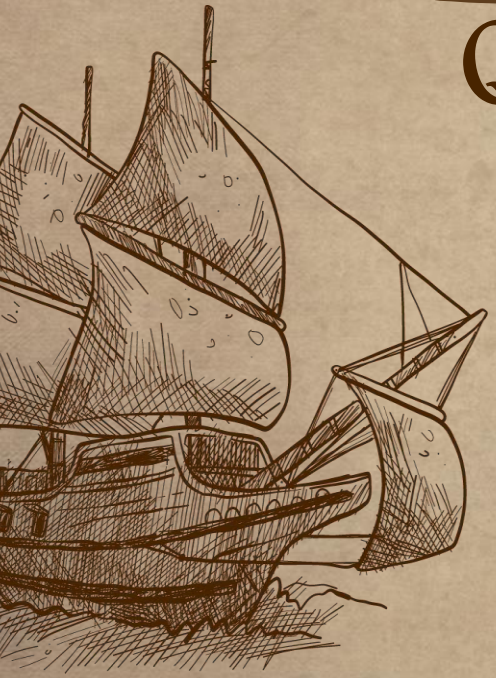
Nile University
جامعة النيل

HistoryQuest: Arabic Question Answering in Egyptian History



Under supervision of :
Dr. Wael Gomaa
ID: 203

Team Members

Samaa Maged

Team Members

Asmaa ElMaghraby

Team Members

Ali Marzban






Team Members

Esraa Negm



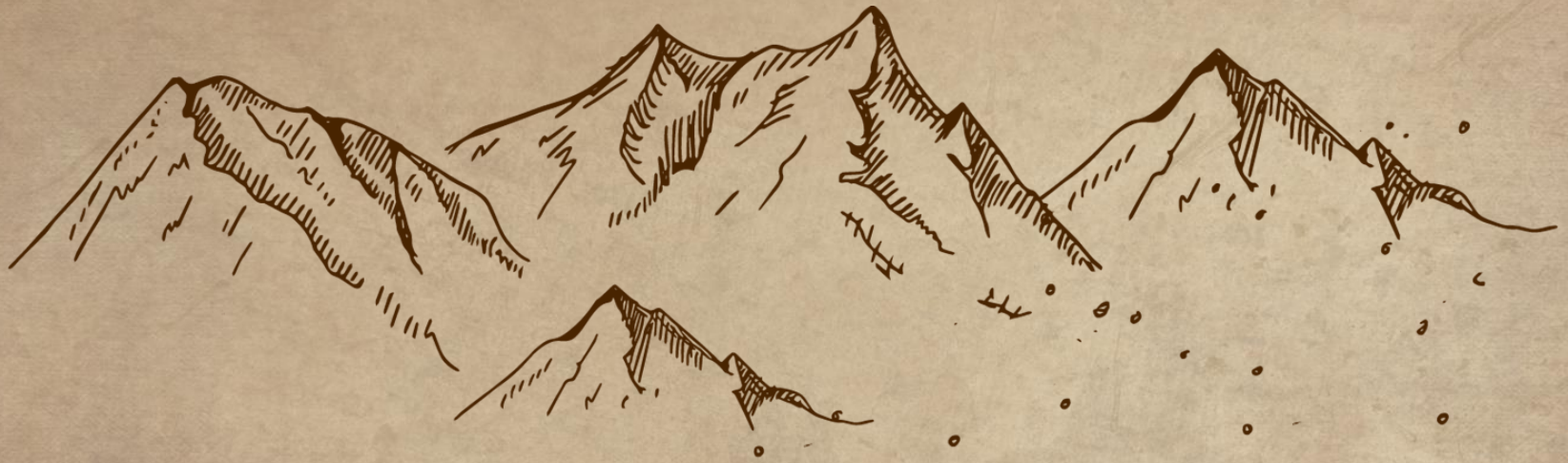


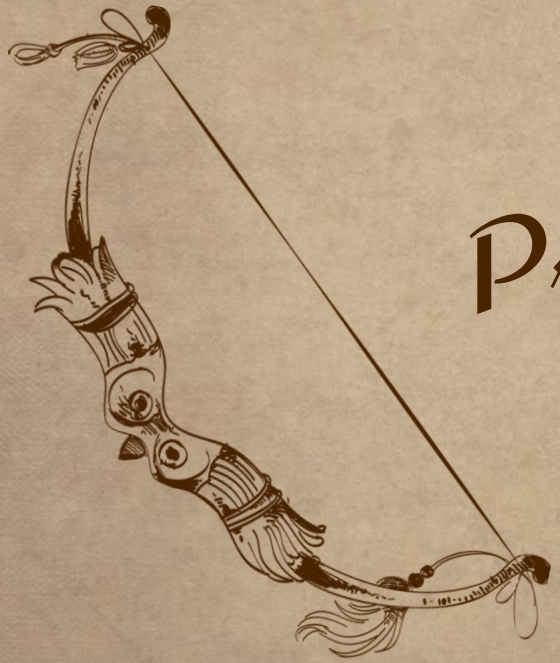

Team Members

Mohamed Essawey

Team Members

Amira Elsharaby



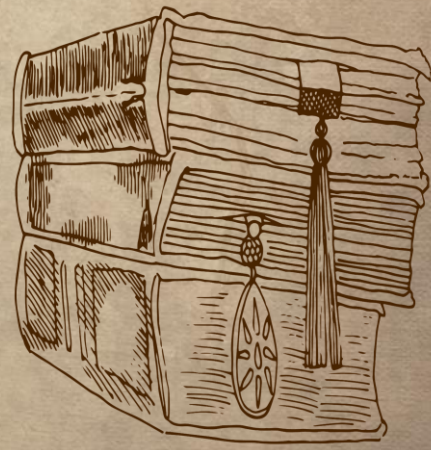


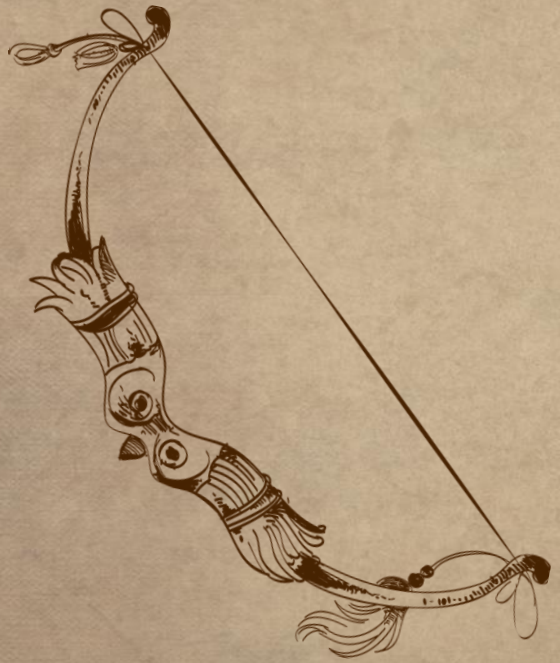
01

Problem defintion

Our project explores how large language models and transformers can improve the way we answer questions about Egyptian history in Arabic.

Using models like AraBERTv2 and LLaMa-3. **This work shows how modern technology can make understanding Egypt's rich history easier and more accessible.**





02

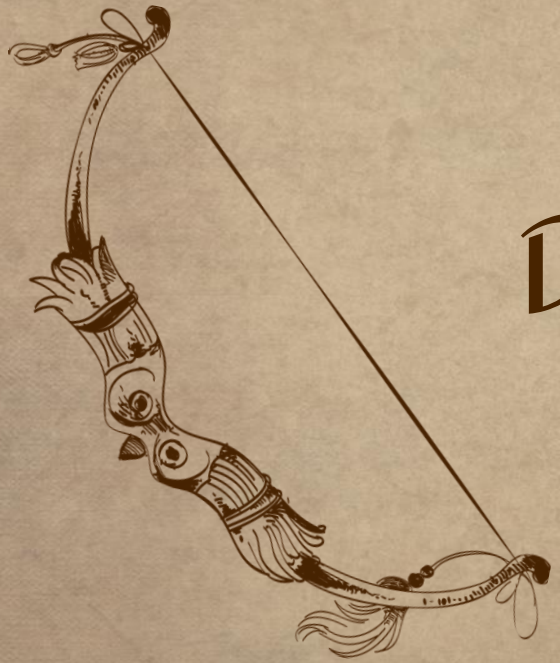
Related Work

Related Work (English QA)

	Dataset Used	Results
Yang and Ishfaq	Stanford Question Answering Dataset (SQuAD) with 100,000+ questions from Wikipedia articles	Achieved 62.23% F1 score and 48.72% EM score on the test set.
Vold and Conrad	PrivacyQA dataset with 1750 questions about privacy policies of mobile applications and over 3500 expert annotation	RoBERTa achieved a 31% improvement in F1-score and a 41% improvement in mean reciprocal rank over traditional SVM.
Izacard and Grave	TriviaQA dataset with over 650,000 question-answer-evidence triples and 95,000 question-answer pairs	Not explicitly mentioned in the paper
Izacard et al.	MMLU, KILT, and NaturalQuestions	Atlas achieved over 42% accuracy on natural questions using only 64 examples,
Wang et al	Four open-domain QA tasks (specific datasets not mentioned)	Proposed improvements in architecture and training led to better utilization of external knowledge. Specific results were not detailed in the provided text

Related Work (Arabic QA)

	Dataset Used	Results
Mozannar et al	Arabic Reading Comprehension Dataset (ARCD) with 1,395 questions from Wikipedia articles.	61.3% F1 score and 90.0% sentence match on ARCD 27.6% F1 score on an open domain version of ARCD.
Alsubhi et al.	AQAD: Large-sized high-quality dataset with 17,000+ questions and answers.	Comparative performance evaluation of the models on the datasets (specific results not detailed in provided)
Atef et al	AQAD	33 Exact Match (EM) and 37 F1-score using mBERT. 32 EM and 32 F1-score using BiDAF model.
Antoun et al	Various Arabic NLP tasks including reading comprehension	Model performance served as a measure of reading comprehension and language understanding capabilities (specific results not detailed in provided text,



03

Data Collection

THE TWO DATASETS

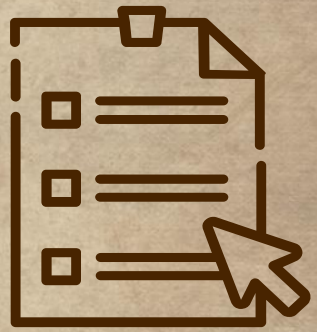
01

Arabic History-QA

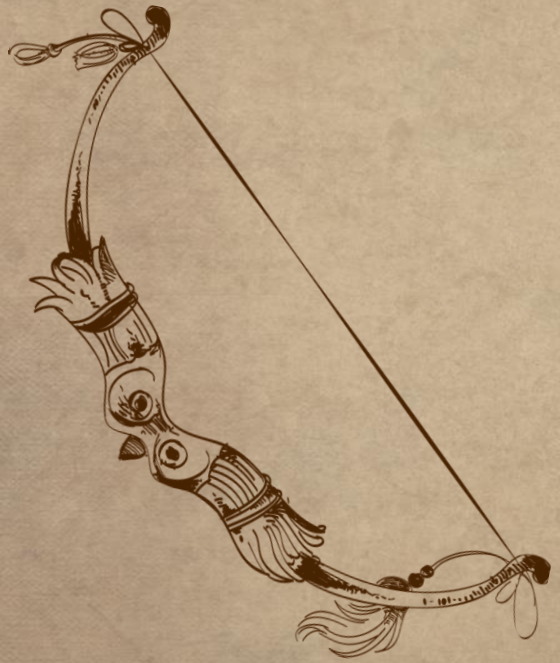
A dataset of 420 question-answer pairs covering key periods in Egyptian history, including the Pharaohs' era, the Roman occupation, the Ptolemaic Dynasty, and the Greek era with Alexander the Great.

02

Contextual Articles Dataset



A collection of web-scraped articles in PDF format, providing a knowledge base for the Retrieval-Augmented Generation (RAG) component of the question-answering system.



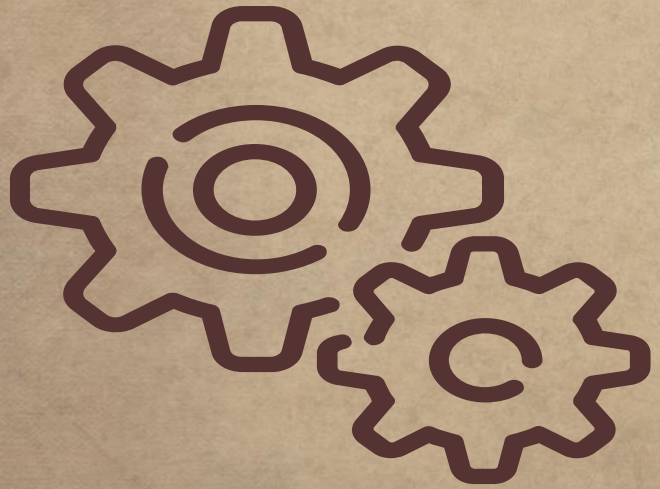
04

Methodology

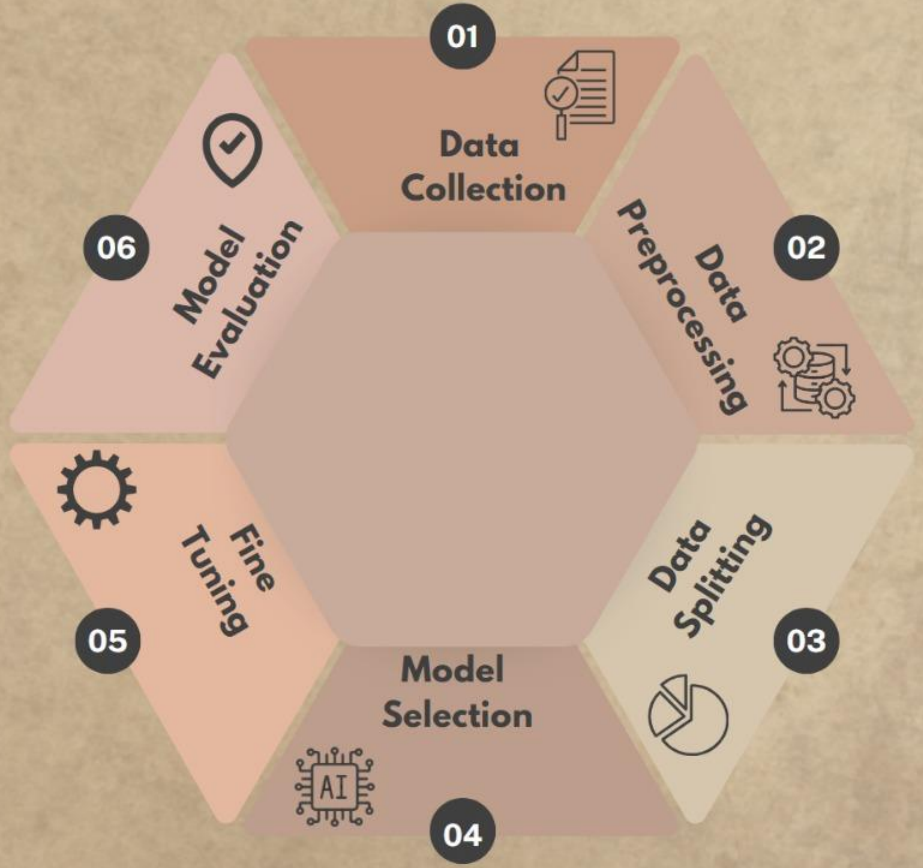
1-Arabic History-QA Methodology

2 - Contextual Articles Dataset Methodology

Arabic History-QA Methodology



Our Pipeline



Data Splitting

95%

399 queries for training

5%

21 queries for testing

Model Selection

(1)AraBERTV2

AraBERTv2, a specialized version of the BERT model for Arabic, Training AraBERTv2 involves 10 epochs with a batch size of 16, adjusting hyperparameters for optimal performance.

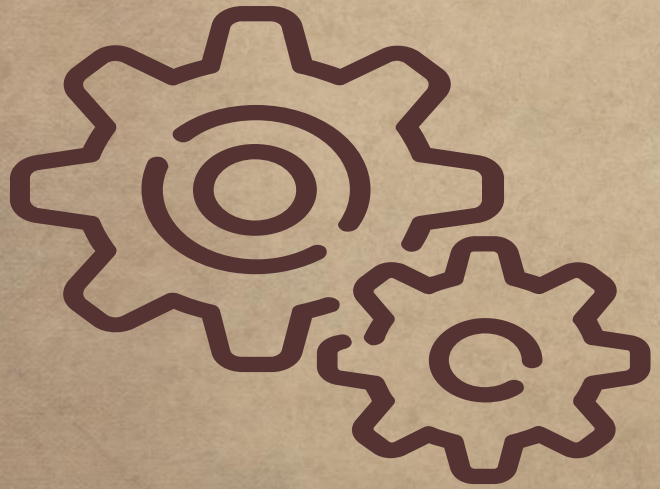
(2)LLaMa-2

LLaMA-2, was tweaked for historical text analysis on the dataset. It underwent 10 epochs of fine-tuning with adjustments, as optimizing LoRA parameters and regulating the learning rate. A LoRA dropout of 0.1 was used to prevent overfitting

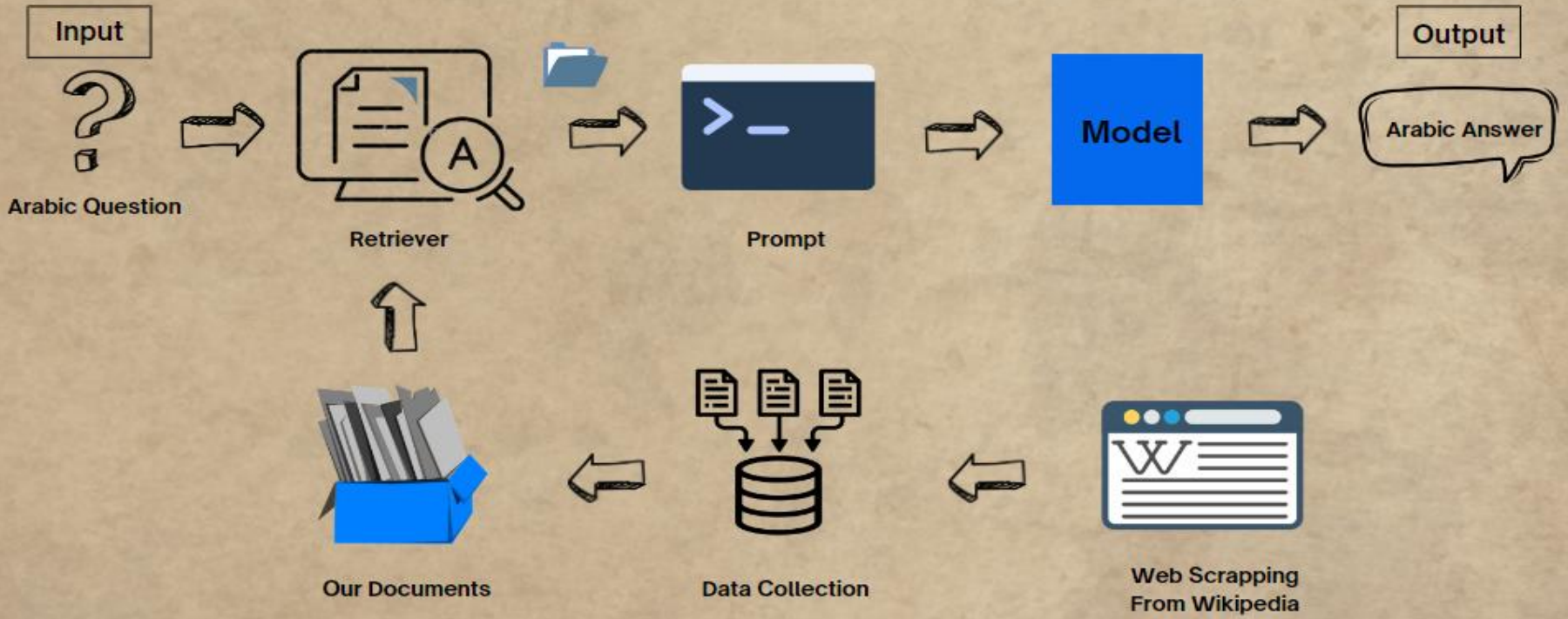
(3)LLaMA-3

Using LLaMA-3 8B parameters for zero-shot learning, use its strong pre-trained architecture. We incorporate LoRA upgrades, with a LoRA Dropout of 0.1 to guarantee the model's robust generalization We set LoRA to 64 for enhanced self-attention and LoRA Alpha to 16 for optimized learning rate management. By improving its efficiency

Contextual Articles Dataset Methodology



The RAG Model processing Pipeline



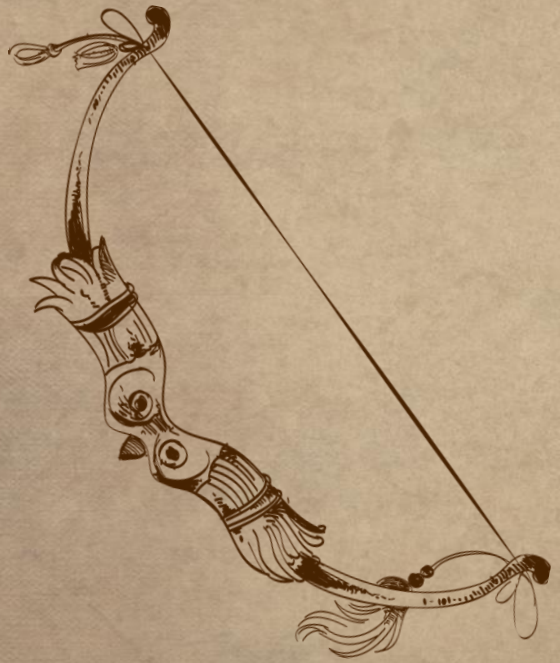
Model Selection

(1)BERT-large-Arabic with RAG

- Indexing historical articles dataset
- Splitting and cleaning the articles into manageable chucks
- Converting these chunks into dense vectors representations using an embedding model (BERT-large-Arabic)
- Retrieving relevant chunks
- Processing the retrieved chunks to generate accurate and contextually relevant responses
- Ensuring the system produces fact-based answers that reflects the complexity of the topics addressed

(2)LLaMa-3 with RAG

- Using the Meta LLaMA-3 model into advanced zero-shot learning capabilities
- Indexing and processing the historical articles
- Splitting articles into manageable chunks
- Embedding chunks into dense vector using BERT-large-Arabic.
- Employing FAISS for efficient similarity search
- Retrieving historical content for informed responses
- Boosting model efficiency with techniques such as bits and bytes quantization



05

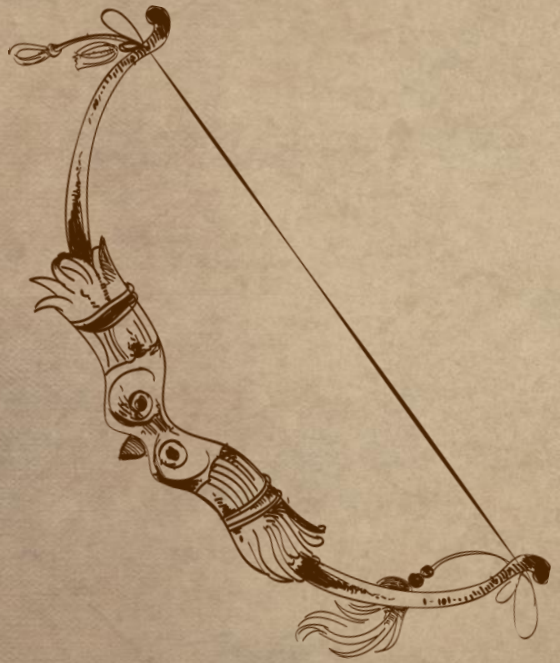
Results

RESULTS

	Dataset	Loss	Correct Predictions	Incorrect Predictions
LLaMa-2	History-QA	0.2202	17 queries	4 queries
LLaMa-3	—	—	13 queries	8 queries
LLaMa-3 with RAG	Contextual Articles Dataset	—	18 queries	3 queries
BERT-large-Arabic with RAG	Contextual Articles Dataset	—	10 queries	11 queries
AraBERTV2	History-QA	—	7 queries	14 queries

Samples

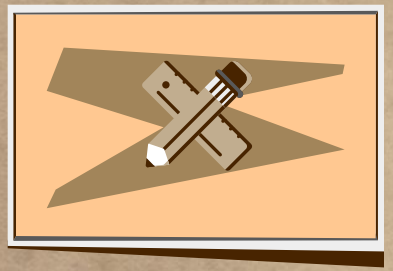
Question	Response
متى حكمت حتشبسوت مصر؟	حكمت حتشبسوت مصر حوالي عام 1458-1479 قبل الميلاد .
أين دفنت حتشبسوت ؟	في وادي الملوك في الحضارة المصرية القديمة .
من قام ببناء مكتبة الإسكندرية ؟	بطلميس الثاني .
ما هو الغرض من بناء هرم خوفو؟	كان الغرض الرئيسي من بناء الهرم هو استخدامه كمقبرة للفرعون



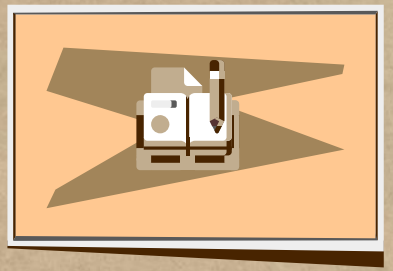
06

Future work

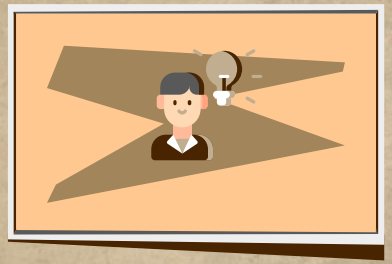
Futre Work



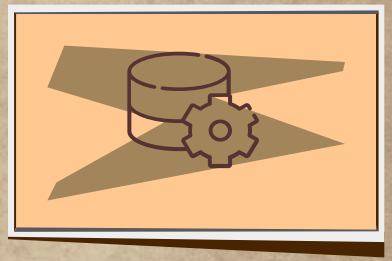
Applying transfer learning approaches to handle multilingual historical documents



Strengthen the generalization capabilities of models across diverse contexts and domains



Refine and fine-tune LLaMa-3 using reinforcement learning with human feedback



Expand the diversity and scale of datasets used for training and evaluation

Thank you