Candidate-Aware Retrieval and Reranking for Multiple-Choice Question Answering: Arabic as a Case Study

Yassine Bouziane
College of Engineering and Architecture
International University of Rabat
Rabat, Morocco
yassine.bouziane@uir.ac.ma

Youness Moukafih
College of Engineering and Architecture
International University of Rabat
Rabat, Morocco
youness.moukafih@uir.ac.ma

Mounir Ghogho
College of Computing
Mohammed VI Polytechnic University
Ben Guerir, Morocco
mounir.ghogho@um6p.ma

Abstract

Large language models (LLMs) have recently achieved impressive results on multiple-choice question answering (MCQA), with retrieval-augmented generation (RAG) emerging as an effective strategy for improving the performance of smaller models. However, existing RAG formulations face persistent challenges: retrieving too many passages often introduces noise, and even when relevant content is retrieved, models may still struggle with partially relevant or conflicting information. Moreover, while LLMs perform strongly on English benchmarks, their accuracy declines substantially on Arabic multi-task evaluations, revealing ongoing issues in cross-lingual transfer and domain adaptation. In this paper, we propose a novel approach, using Arabic as a representative case study, that jointly models the relevance of both the question and its candidate answers when selecting contextual passages. The method employs a lightweight reranker trained with a hybrid regression–triplet loss objective to identify passages that provide discriminative and reliable evidence. Extensive experiments across multiple model sizes and humanities domains show that our approach consistently outperforms both standard RAG baselines and reranker baselines, delivering two- to threefold improvements while remaining competitive with considerably larger models.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across diverse Natural Language Processing (NLP) tasks, including text classification, summarization, reasoning, and dialogue generation. This success is largely attributed to large-scale pretraining and instruction tuning, which endow LLMs with strong generalization in zero-shot and few-shot settings. Among these tasks, Multiple-Choice Question Answering (MCQA) has emerged as a key benchmark, capturing model's comprehension, reasoning, and factual retrieval abilities.

Despite remarkable advances in English, LLMs continue to underperform in Arabic due to enduring gaps in data coverage, linguistic transfer, and evaluation design. Results from ArabicMMLU (2024) show significant accuracy declines on Arabic multi-task benchmarks compared with their English counterparts, revealing incomplete cross-lingual transfer and limited transparency in training data, even for leading models such as GPT-4. Beyond general knowledge benchmarks, Arabic MCQA tasks expose deeper deficiencies. Studies like AraSTEM demonstrate that domain-specific questions exacerbate errors linked to Arabic morphology, specialized terminology, and distractor sensitivity, despite extensive multilingual pretraining. These limitations underscore the need for Arabic-centered benchmarks and modeling strategies that account for the language's structural and semantic features.

We address these challenges through a novel RAG formulation for MCQA that explicitly integrates answer candidates into the retrieval and ranking process. Our method combines efficient evidence selection with targeted reranking to better align retrieved content with the most plausible answers. We use Arabic as a case study. Specifically,

Findings of the Association for Computational Linguistics: ACL 2026, pages 8967–8977

our main contributions are as follows:

- *Answer candidate-aware evidence selection.* Retrieval and reranking operate on triplets (q, a_i, c), where q denotes the question, a_i represents the i-th candidate answer, and c is a candidate passage from the corpus. Unlike question-only retrieval, this formulation explicitly models the interaction between each answer option and its supporting context, prioritizing passages that either substantiate or refute individual candidates.

- *Disentangling informative and resolving context.* A lightweight reranker trained with a joint regression and triplet loss distinguishes between context that merely provides background information and context that directly resolves the answer. This separation mitigates common "relevant-but-insufficient" failure modes.

- *Corpora and supervision.* We curate HUMCORPUS, a collection of Arabic educational and general-knowledge materials, to serve as the retrieval base. We also construct RANKMCQ, a dataset with controlled positive, partially positive, and negative passages designed to supervise evidence–answer alignment.

- *Resource-aware retrieval.* A managed Milvus-based vector database decouples embedding storage and similarity search from local computation. This design enables reproducible deployment in resource-constrained environments and allows simultaneous use of the vector database and reranker on the same local resources.

- *Evaluation protocol.* We extend the *ArabicMMLU* prompt with evidence concatenation to evaluate candidate-aware retrieval and reranking in Arabic MCQA, enabling smaller models to match or surpass larger ones for low-resource languages, marking a step forward for such languages.

2 Related Work

Dedicated Arabic LLMs, such as Jais, AceGPT, and MARBERT, have intensified the demand for *authentic* Arabic evaluations that move beyond translated benchmarks. Recent efforts, including *ArabicMMLU* (Koto et al., 2024), *AraSTEM* (Mustapha et al., 2024), and *ArabicaQA* (Abdallah et al., 2024), emphasize native, context-rich questions covering both general knowledge and curricular domains, thereby addressing limitations inherent in direct translations. In parallel, retrieval-augmented generation (RAG) has emerged as a practical approach for linking model predictions to external knowledge sources (Karpukhin et al., 2020; Edge et al., 2024). However, two practical challenges exists. First, increasing the number of retrieved passages (k) often adds noise and yields diminishing returns, with performance typically stabilizing around $k \approx 10$ (Yu et al., 2024a). Second, even when relevant evidence is retrieved, models often fail to extract the correct answer among contradictory or partially relevant passages (Barnett et al., 2024). To address these limitations, recent research has further refined this process through advances in ranking and evidence selection combining dense retrieval, reranking, and filtering techniques (Yu et al., 2024b,a; Ram et al., 2023). Within MCQA pipelines, earlier studies primarily used retrieval to support *question generation* and *distractor construction* (Pawar and Makwana, 2022; Zhu et al., 2024; Dhole and Manning, 2020), while more recent work integrates contextual grounding to improve answer selection (N et al., 2025; Chen et al., 2025).

We argue that retrieval and ranking should be conditioned not only on the question but also on the candidate answers. Several methods already exploit interactions between questions and answer options in retrieval or evidence scoring. (Yadav et al., 2019a) uses BM25 to retrieve supporting paragraphs and scores candidates via alignment of *question+answer* with paragraphs using heterogeneous embeddings. (Singh and Shrivastava, 2025) fine-tunes a retriever to approximate the concatenated query (question plus options) for long-context MCQA. Both focus on retrieval, with ranking based directly on retrieval or alignment scores, not on a separate reranking stage.

In the other hand, (Yadav et al., 2019b) includes reranking, selecting justification sets using relevance, overlap, and coverage scores. However, it is unsupervised, operates at the set level, and does not condition on individual answer options or use explicit supervision signals for answer-evidence alignment. Our reranker, by contrast, is trained

on RANKMCQ with positive, partially positive, and negative supervision, operates at the passage level, and is conditioned on each answer candidate, enabling more precise, interpretable, and answer-aware ranking for Arabic MCQA. This allows our system to not only retrieve relevant passages but also to identify those that are decisive for each candidate answer, which is a significant step beyond prior work.

3 Method

To address the problems outlined in the previous section, we present a four-step method, illustrated in Figure 1, that aims to improve how LLMs perform on MCQA tasks. In this approach, a reranker is trained to select passages using both the question and its answer candidates as input. This joint consideration helps ensure that the retrieved information is directly relevant for identifying the correct answer, while minimizing the inclusion of context that is only loosely related to the question or the answer options.

3.1 Step-by-Step Method Description

In **Step (1)**, we employ Zilliz Cloud, a managed vector database built on Milvus® (Wang et al., 2021), as our primary retrieval engine. Because the experimental setup operates under limited computational resources, we adopt this cloud-based solution to offload embedding storage, indexing, and similarity search operations. We then define the following RAG configuration: given a question q and a set of n answer choices $A = \{a_1, a_2, \ldots, a_n\}$, we construct candidate queries by concatenating each answer choice a_i with the question.

In **Step (2)**, we retrieve the top-k candidate contexts $c(q, a_i)$ for each query using cosine similarity as the retrieval metric. In contrast, a standard RAG system typically retrieves passages based solely on the question and selects the highest-scoring passage R as the retrieval result used to condition the generator.

In **Step (3)**, our objective is to evaluate the joint relevance between the question, its answer choices, and the retrieved contexts. We hypothesize that lower-ranked passages may, in some cases, contain more accurate or contextually useful information than R, depending on their relevance to all candidate answers. To this end, we train a reranker that assigns a relevance score to each of the k candidate contexts $c(q, a_i)$ for each answer choice a_i, ranking how well each context supports the question in relation to that choice. The context with the highest score is then selected as the most relevant for a_i. Repeating this process for all answer choices yields n matched contexts $S = \{s_1, s_2, \ldots, s_n\}$, which are subsequently provided to the LLM as supporting information for answer prediction in **Step (4)**.

For consistency and fair comparison, we follow the *ArabicMMLU* prompt template illustrated in Figure 2, extending it to include the retrieved context component. Given the question `[question]` and all answer options `[options]`, the placeholder `[context]` corresponds to the retrieved contexts S from **Step (3)**, where all matched contexts $s_1, s_2, \ldots, s_n$ are concatenated into a single input sequence.

Subject	Country	# Chunks
Islamic Studies	Palestine - Jordan	257 084
Philosophy	Egypt	8 463
Law	Morocco	16 806
History	Jordan - Egypt	49 833

Table 1: HumCorpus chunks distribution and country source of information. Note that the number of chunks varies significantly between subjects, highlighting the difficulty of collecting large, representative samples for some topics within certain countries.

3.2 HumCorpus Dataset

Due to the wide range of fields, subjects, and educational levels covered in the ArabicMMLU benchmark, we focus our evaluation primarily on the humanities domain, targeting the most challenging difficulty levels (High School and University). We construct a new dataset named HumCorpus[1], which serves as the knowledge source for our RAG system. HumCorpus is composed of publicly available Arabic educational textbooks and documents, with the majority of materials collected from resources within the Arabic educational system to ensure strong alignment with the benchmark content. Additionally, we include general knowledge documents related to the same subjects to broaden coverage and enhance the comprehensiveness of the knowledge base.

Table 1 shows the details for each subject of HumCorpus. To represent textual information se-

[1] https://github.com/TheSoloLeveling/HumCorpus-RAGNARank

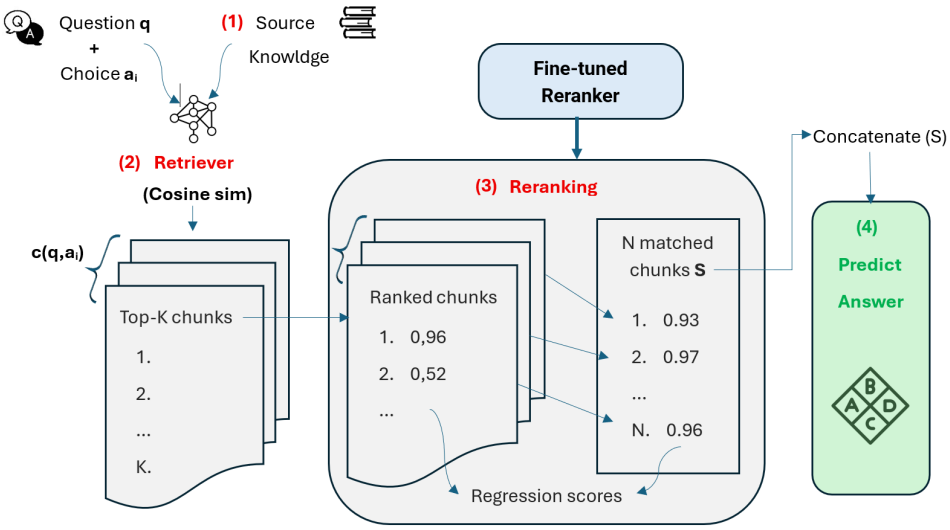


Figure 1: Workflow of our method for MCQA integrating retrieval and reranking.

This is a **[SUBJECT]** question for **[LEVEL]** in **[COUNTRY]**.
Select the correct answer given the context **[context]**.

Question: **[question]**

[options]

Answer:

This is a **[SUBJECT]** question for **[LEVEL]** in **[COUNTRY]**.
Select the correct answer.

Question: **[question]**

[options]

Answer:

Figure 2: Right is the template prompt used in our method and left is the ArabicMMLU template prompt.

mantically, the data is first divided into manageable chunks of up to 1,000 tokens, with a 200-token overlap between consecutive segments. Each chunk is then encoded using an Arabic Sentence-BERT model (Reimers and Gurevych, 2019), generating 768-dimensional embedding vectors that are indexed in the Zilliz Cloud vector database for semantic similarity search.

3.3 RankMCQ Dataset

To train our reranker, we need ideally three distinct types of context to be able to score effectively a question-context-answer relevance: (i) a *positive context*, which contains accurate information directly leading to the correct answer; (ii) a *partially positive context*, consisting of factual and relevant information about the answer but inadequate to fully resolve the question; and (iii) a *negative context*, which includes false or misleading information to both question and answer. Since no publicly available dataset adheres to this specific design, we created RankMCQ[2], a new training/validation dataset leveraging the PalmX 2025 Islamic Culture dataset (Alwajih et al., 2025).

RankMCQ was constructed through a two-stage pipeline combining LLM-based generation with human verification. Positive and partially positive contexts were generated using the prompts illustrated in Figure 3. Each generated context was subsequently reviewed by human annotators alongside the corresponding question and answer candidates according to three criteria: (i) coherence and linguistic adequacy; (ii) candidate-specificity; and (iii) label correctness, confirming that the partially positive passages provide useful background knowledge without resolving the question. This is crucial, as it enables the reranker to learn to assign appropriate rankings to close contexts. Overall, manual intervention was limited: the majority of generated contexts were accepted without modification, with edits or regeneration required only for a small subset of particularly complex or ambiguous samples.

All contexts are generated using LLAMA4 SCOUT 17B INSTRUCT which is available through

[2] https://huggingface.co/datasets/TarnishedNathe/islamic-QCM

meta API [3]

3.4 Answer-Question Context Reranker Training

We build a reranking model restricted to a small dataset, using a combined regression and triplet loss approach on the RankMCQ dataset with contextual passages. Each sample of RankMCQ consist of a question q, the correct answer a^T where $a^T \subseteq A$ and 3 different types of context components: positive c^+, partially positive c^0 and negative c^-.

i. Input Representation: We start by generating multiple training examples per sample to capture different context-answer relationships. Each training example is formed by concatenating the following components, separated by a special token [SEP]:

1. q [SEP] c^+ [SEP] a^T: we pair the question and the true answer with the positive context which encourages the model to learn how the correct answer is supported by evidence from an informative context passage. Therefore, we assign the highest possible regression score of 1.

2. q [SEP] c^0 [SEP] a^T: here the question and the true answer is paired with the partially positive context. A medium score of 0.7 is assigned to indicate partial relevance, meaning the context is highly supportive with respect to the answer, but also does not give robust evidence for the question.

3. q [SEP] c^- [SEP] a^T: negatives examples are assigned with the lowest score of 0, where we pair the question and the true answer with the negative context. We strongly penalize such cases in its ranking predictions.

To enhance the efficiency of our reranker in the context of limited training data, we augment more hard negatives (partially positive contexts) examples with the following sentence inputs:

4. r_q [SEP] c^+ [SEP] a^T: r_q is a randomly selected question from the dataset (not the one associated with a^T), paired with the positive context and the true answer. This is another reflection of the partial relevance with a score of 0.7, where context-answer pairs matches, but do not match the question.

5. q [SEP] c^+ [SEP] a^R: The question and context matches, but paired with incorrect random answer a^R chosen from the same dataset, which we assign a score of 0.

ii. Model Architecture: These input sequences are passed through a pre-trained Transformer encoder to obtain dense representations. We define the encoder function $\mathbf{h}(x)$ where x is the input token sequence, of length L, and $\mathbf{h}(x) \in \mathbb{R}^{d \times L}$ is the sequence of hidden states, with L tokens and hidden size d. Then, the pooled representation $\mathbf{z} \in \mathbb{R}^{d \times 1}$ is obtained by extracting the hidden state corresponding to a learnable `[CLS]` token $\mathbf{z_x} = \mathbf{h}(x)[0]$ (i.e. first column of $\mathbf{h}(x)$). This pooling approach is suitable as $\mathbf{z}$ acts as a global summary embedding of the entire input sequence.

Joint Embedding Encoder: We train the encoder jointly using triplet loss and regression to separate positive from partially positive contexts. The anchor, positive, and negative inputs share the same encoder with $\mathbf{x}$ to produce embeddings $\mathbf{z}_a, \mathbf{z}_p, \mathbf{z}_n \in \mathbb{R}^d$ for triplet loss computation (3). The anchor is formatted as q [SEP] a^T, with positive as c^+ and and negative as c^0.

Regression Head: A feed-forward layer projects the pooled embedding $\mathbf{z_x}$ to a scalar regression score $\hat{y}$ indicating the relevance or correctness of the input:

$$\hat{y} = \sigma\big(\mathbf{w}\mathbf{z_x} + b\big) \tag{1}$$

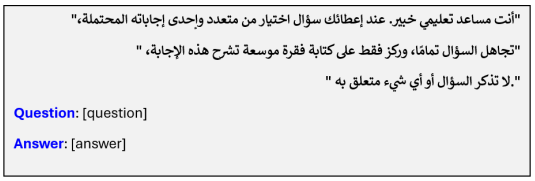
where $\mathbf{w} \in \mathbb{R}^{1 \times d}$ and $b \in \mathbb{R}$ are learned parameters, and $\sigma(\cdot)$ denotes the sigmoid function ensuring $\hat{y} \in [0, 1]$.

iii. Training: we optimize the reranker model with a combination of two losses: a regression loss and a triplet loss. Each loss operates on specific inputs as described above.

Although our relevance labels fall into three discrete categories $(0, 0.7, 1)$, they exhibit an inherent order and semantic continuity rather than representing independent classes. Therefore, instead of treating the problem as categorical classification, we cast it as a form of ordinal regression (Cheng et al., 2008), (Shi et al., 2023), where the target values correspond to progressively higher degrees of contextual relevance.

The regression loss supervises the predicted relevance score $\hat{y}$ against the ground-truth soft target score y using Mean Squared Error (MSE):

[3] https://ai.meta.com/llama/



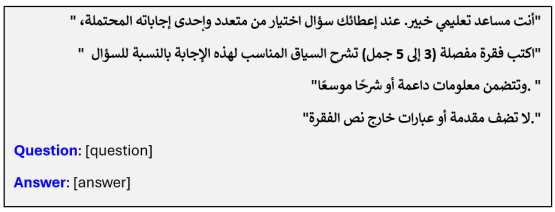


Figure 3: Left is the template prompt used to generate partially positive context and right is the template prompt used to generate positive context.

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2}$$

where $\hat{y}_i \in [0,1]$ is the predicted regression score for the i-th training example, produced by the regression head; $y_i \in [0,1]$ is the ground-truth soft label indicating the relevance or correctness of the i-th input example. (0 for negative context, 0.7 for partially positive context, 1 for positive context); and N is The number of examples in the batch.

The triplet loss uses cosine similarity-based distances to encourage the embedding of an anchor input $\mathbf{z}_a$ to be closer to positive $\mathbf{z}_p$ than to negative $\mathbf{z}_n$ embeddings by a margin m. The triplet loss over M valid triplets is:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{M}\sum_{j=1}^{M}\max\Big(0,\, d(\mathbf{z}_a^j, \mathbf{z}_p^j) - d(\mathbf{z}_a^j, \mathbf{z}_n^j) + m\Big) \tag{3}$$

where the distance function used for triplet loss between 2 vectors u and v is:

$$d(\mathbf{u},\mathbf{v}) = 1 - \text{cosine\_sim}(\mathbf{u},\mathbf{v}) \tag{4}$$

The total training loss combines the two components with a weighting factor $\alpha \in [0,1]$:

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{triplet}} + (1-\alpha)\,\mathcal{L}_{\text{reg}} \tag{5}$$

4 Experiments

In this section, we compare our method against the standard RAG retrieval method (named in Section 3 as R) and baseline models on ArabicMMLU, highlighting its superior zero-shot performance and robustness in low-resource settings. All RAG hyperparameters—chunk size (1,000 tokens), chunk overlap (200 tokens), top-k (k=10), embedding dimension (768), and cosine-similarity retrieval—are fixed across all experimental variants to ensure a controlled and fair evaluation. It is worth noting that the domain alignment observed in HumCorpus is an inherent characteristic of RAG-based systems rather than a property introduced by our method. Since the retrieval configuration is held constant across all experimental variants, any performance differences observed between systems arise directly from the reranking stage, the component where our candidate-aware method operates.

4.1 Experiment Setup

Datasets. We consider four evaluation datasets from the ArabicMMLU humanities domain: Islamic Studies (HS) with 377 questions, Philosophy (HS) with 42 questions, Law (U) with 317 questions, and History (HS) with 763 questions. The labels "HS" and "U" denote High School and University levels, respectively.

Baselines. We focus on models with fewer than 7 billion parameters to prove strong performance under limited-resource conditions. Our evaluation includes three baseline families, each with multiple sizes: Qwen3 (4B and 1.7B) (Yang et al., 2025), Llama-3.2 (3B and 1B) (Dubey et al., 2024), and Mistral (7B and 7B-Instruct) (Jiang et al., 2023). For comparative analysis, we also include five open-source models from the ArabicMMLU benchmark: XGLM (Lin et al., 2022), LLaMA2 (Touvron et al., 2023), AceGPT (Huang et al., 2023), BLOOMZ (Muennighoff et al., 2022), and Jais (Sengupta et al., 2023) across different model sizes, alongside two closed-source models, GPT-3.5 (gpt-3.5-turbo) (Ouyang et al., 2022) and GPT-4 (gpt-4-0613) (Achiam et al., 2023).

Set-Up. Following (Koto et al., 2024), we determine the answer based on the highest probability among all possible options, and we conduct all experiments using the prompts in Figure 2. Our evaluation considers three settings: baseline without retrieval, with standard RAG, and with our pro-

posed reranking approach. We set $k = 10$ to select the top-10 retrieved contexts for both the standard RAG retrieval and our reranker. For ArabicMMLU baselines, we exclusively report the official results as released by the benchmark authors.

Evaluation Metrics. Following (Koto et al., 2024), we use `accuracy` as the metric for all experiments.

4.2 Reranker Implementation

We leverage an openly available Arabic reranker[4], which is fine-tuned for paragraph ranking from a state-of-the-art Arabic language embedding model (Nacar et al., 2025), as the backbone for our reranker model. We further fine-tune this model on the newly created RankMCQ dataset following the approach detailed in Subsection 3.4. Training is conducted with a batch size of 8 for 20 epochs using the Adam optimizer (Kingma, 2014), a learning rate of $2e-5$, and a weighting factor $\alpha = 0.2$. The resulting model effectively ranks triplet questions, contexts, and answer relevance, enhancing the accuracy of our Arabic RAG pipeline.

5 Results

This section presents the experimental findings, highlighting the comparative performance of the proposed method against ArabicMMLU baselines and standard RAG configurations across multiple backbone models and subject domains.

5.1 Main results

As shown in Table 2 and Table 3, our method outperforms baseline reranker, standard RAG retrieval and baseline configurations when averaging across the full benchmark. The results show that we consistently obtain best results across nearly all evaluated backbone models. Notably, our method with MISTRAL-7B as backbone achieves the highest average score of 47.3, showing significant gains over the baseline (31.3), standard RAG (42.5) and baseline reranker (41.8). This trend persists for other models such as QWEN3-1.7B and QWEN3-4B, where we observe robust improvements, particularly on complex domains like Philosophy and Law.

Our method demonstrates more Consistency. we observe that the impact of incorporating retrieved context varies significantly between models across different subjects, reflecting differing abilities to leverage additional information. For instance, in the Law subject, the baseline reranker leads to a 2-point drop in accuracy for QWEN3-4B, and standard RAG retrieval further decreases performance by 5 points. The degradation is even more pronounced with LLAMA-3.2-1B, where standard RAG causes a loss of over 13 points and the baseline reranker results in a 5-point drop. In contrast, our method achieves a 5-point improvement under the same conditions. Although the baseline reranker slightly outperforms our approach in a few isolated cases, these gains are marginal. On average, our method still yields the highest overall improvement.

These observations suggest that certain architectures may struggle to integrate external context effectively, sometimes resulting in harmful interference or distraction, which limits reranker's adaptability. In contrast, our method demonstrates more consistent performance, with far fewer cases of performance degradation; on most models, it either improves or maintains accuracy, highlighting its robustness and reliability in leveraging external information.

5.2 Performance against ArabicMMLU baselines

The comparative results presented in Table 4 clearly highlight the strong competitiveness of the proposed method against the large-scale ArabicMMLU baselines. In particular, the Qwen3-4B variant consistently surpasses several well-established Arabic language models such as BLOOMZ and AceGPT across all evaluated domains. The improvement is especially pronounced in the Law subject, where Qwen3-4B achieves a remarkable gain of 15 points over BLOOMZ (7B) and 21 points over AceGPT-chat (13B) in terms of accuracy. Even more notably, when compared to BLOOMZ in the Philosophy domain, Qwen3-4B demonstrates a substantial 31-point advantage, underscoring the model's enhanced reasoning and knowledge grounding capabilities in complex, textually rich disciplines.

In comparison with much larger and more powerful systems, including Jais-chat (30B), GPT-3.5 (175B), and GPT-4, our method maintains highly competitive performance despite its smaller scale. The Qwen3-4B model, for instance, achieves nearly identical results to GPT-4 in Philosophy (74.3 vs. 74.4 accuracy) while outperforming GPT-3.5 and Jais-chat by 15 and 8 points, respectively. These

[4] https://huggingface.co/oddadmix/arabic-reranker

Subject	Method	Qwen3-4B	Qwen3-1.7B	Mistral-7B	Mistral-7B-instruct	Llama-3.2-3B	Llama-3.2-1B
All Subjects	our method	**57.5**	**45.8**	**47.3**	**45.3**	48.0	**27.9**
	baseline reranker	56.4	40.6	41.8	39.4	**49.8**	26.3
	standard RAG	55.2	42.9	42.5	40.0	47.3	22.7
	baseline	53.7	38.5	31.3	37.3	42.8	26.2

Table 2: Zero-shot average performance across all subjects.

Subject	Method	Qwen3-4B	Qwen3-1.7B	Mistral-7B	Mistral-7B-instruct	Llama-3.2-3B	Llama-3.2-1B
Philosophy (HS)	our method	74.3	**59.0**	**61.5**	**56.4**	58.9	**43.5**
	baseline reranker	**76.3**	51.2	43.5	48.7	**69.2**	33.3
	standard RAG	71.2	56.4	43.5	51.2	55.4	25.6
	baseline	58.9	41.0	20.5	43.5	48.7	38.4
Islamic Studies (HS)	our method	57.7	47.0	50.1	**53.8**	47.6	**29.4**
	baseline reranker	**58.9**	**47.6**	**51.2**	49.1	**51.2**	25.4
	standard RAG	58.9	45.8	51.8	49.9	44.3	22.4
	baseline	57.9	47.3	44.9	43.4	38.9	18.8
Law (U)	our method	56.3	**42.6**	**43.9**	**41.7**	57.6	15.6
	baseline reranker	54.1	39.1	41.0	34.3	50.0	15.6
	standard RAG	51.2	36.6	41.7	35.0	**61.1**	15.2
	baseline	**56.6**	33.7	28.9	36.3	52.5	**17.8**
History (HS)	our method	**41.7**	**34.7**	**33.4**	**29.4**	28.1	24.2
	baseline reranker	36.3	24.4	31.5	25.6	29.0	**31.1**
	standard RAG	39.4	32.8	32.9	24.1	28.0	27.6
	baseline	41.3	31.9	30.7	26.0	**31.3**	30.0

Table 3: Zero-shot performance comparison across different subjects.

Model (#parameters)	Philosophy (HS)	Islamic Studies (HS)	Law (U)	History (HS)
AceGPT-chat (13B)	53.8	51.3	52.7	40.5
BLOOMZ (7B)	59.0	52.8	25.9	38.9
Qwen3* (4B)	**74.3**	**57.7**	**56.3**	**41.7**
Mistral* (7B)	61.5	50.1	43.9	33.4
Llama-3.2* (3B)	58.9	47.6	**57.6**	28.1
Jais-chat (30B)	66.7	66.9	33.1	50.6
GPT-3.5 (175B)	59.0	62.4	55.8	42.7
GPT-4 (NA)	**74.4**	76.7	66.9	54.1

Table 4: Cross-scale reference comparison on ArabicMMLU humanities datasets. Models marked with * ($\leq$7B parameters) are evaluated with our proposed method; large-scale models are included as contextual baselines without our method, following common practice in the RAG literature. Baselines average over all education levels, including the easiest ones, while models leveraging our method uses only high school and university questions, giving reference baselines a relative advantage.

findings demonstrate the efficacy of our reranked retrieval-augmented approach in enabling smaller models to achieve strong performance with minimal computational resources and training cost.

5.3 Ablation on loss terms

We conduct an ablation study to evaluate the contribution of the contrastive loss to the overall model performance. Both variants share the same backbone and evaluation protocol, isolating the effect of the contrastive loss component. Figure 4 and Figure 5 shows that the model with contrastive loss achieves lower loss and faster decrease over time, reaching a lower loss level compared to the regression-only variant. This observation confirms that the contrastive loss increases the separation between classes and improves retrieval accuracy by reducing the variance between positive and partially positive samples.

6 Conclusion

In this paper, we introduce a novel RAG method designed for context ranking to enhance the performance of smaller models on MCQA task. Our ap-

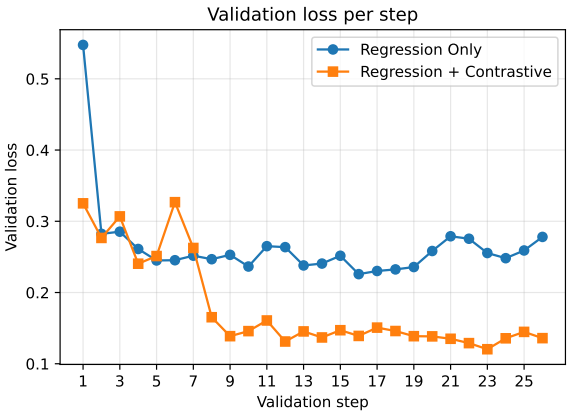


Figure 4: Validation loss for each validation step.

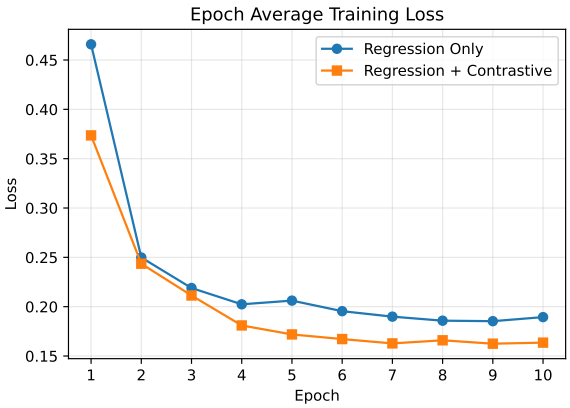


Figure 5: Training loss for each epoch.

proach explicitly considers the relevance of answer choices when selecting the most informative context to effectively solve questions. Experimental results show consistent performance gains across ArabicMMLU humanities benchmarks, outperforming popular Arabic and multilingual LLMs in zero-shot settings and remaining competitive with larger models. While we use Arabic as the primary case study, the proposed pipeline is particularly relevant for low-resource languages facing similar data scarcity challenges; extending this pipeline to additional domains and languages remains an important direction for future work.

Limitations

Our proposed method demonstrates notable improvements over standard RAG retrieval. However, is not without limitations. First, although the reranker effectively filters noisy or partially relevant passages, its performance still depends on the quality of the underlying retriever; if the retrieval stage fails to capture semantically rich or diverse candidates, even a perfect reranker cannot fully compensate. Second, the combined regression and triplet loss objective demands careful data balancing to avoid overfitting to particular relevance patterns or domain-specific distributions. In particular, partially positive and positive contexts can often be very similar or even overlapping, which poses a challenge for the model to effectively distinguish between them during training. Finally, unlike standard retrievers that typically retrieve a single context per question, our approach retrieves N candidate contexts jointly considering question and answer relevance, which can be memory-intensive and computationally costly in low-resource environments. Nonetheless, our method is versatile, allowing integration with various backbone models across different subject domains, while delivering consistent performance improvements.

References

Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fakhraddin Alwajih, Abdellah El Mekki, Hamdy Mubarak, Majd Hawasly, Abubakr Mohamed, and Muhammad Abdul-Mageed. 2025. Palmx 2025: The first shared task on benchmarking llms on arabic and islamic culture. *arXiv preprint arXiv:2509.02550*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199.

Tingwei Chen, Jiayi Chen, Zijian Zhao, Haolong Chen, Liang Zhang, and Guangxu Zhu. 2025. First token probability guided rag for telecom question answering. *arXiv preprint arXiv:2501.06468*.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. 2008. A neural network approach to ordinal regression. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1279–1284. IEEE.

Kaustubh D Dhole and Christopher D Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. *arXiv preprint arXiv:2004.08694*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, and 1 others. 2023. Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ahmad Mustapha, Hadi Al-Khansa, Hadi Al-Mubasher, Aya Mourad, Ranam Hamoud, Hasan El-Husseini, Marwah Al-Sakkaf, and Mariette Awad. 2024. Arastem: A native arabic multiple choice question benchmark for evaluating llms knowledge in stem subjects. *arXiv preprint arXiv:2501.00559*.

Pradeesh N, Remya T, MG Thushara, K Arun Krishna, and Pranav V. 2025. Retrieval-augmented generation for multiple-choice questions and answers generation. *Procedia Computer Science*, 259:504–511. Sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), held in Uttarakhand, India.

Omer Nacar, Anis Koubaa, Serry Sibaee, Yasser Al-Habashi, Adel Ammar, and Wadii Boulila. 2025. Gate: General arabic text embedding for enhanced semantic textual similarity with matryoshka representation learning and hybrid loss training. *Preprint*, arXiv:2505.24581.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Chandrashekhar S Pawar and Ashwin Makwana. 2022. Comparison of bert-base and gpt-3 for marathi text classification. In *Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021*, pages 563–574. Springer.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2023. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3):941–955.

Manish Singh and Manish Shrivastava. 2025. Options-aware dense retrieval for multiple-choice query answering. *Preprint*, arXiv:2501.16111.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, and 1 others. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 international conference on management of data*, pages 2614–2627.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019a. Alignment over heterogeneous embeddings for question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2681–2691, Minneapolis, Minnesota. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019b. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2578–2589, Hong Kong, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 121156–121184. Curran Associates, Inc.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, and 1 others. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.