Open-domain Arabic Conversational Question Answering with Question Rewriting

Mariam E. Hassib **Nagwa El-Makky** **Marwan Torki**
Faculty of Engineering, Alexandria University, Egypt
{mariam.hassib23, nagwamakky, torki}@alexu.edu.eg

Abstract

Conversational question-answering (CQA) plays a crucial role in bridging the gap between human language and machine understanding, enabling more natural and interactive interactions with AI systems. In this work, we present the first results on open-domain Arabic CQA using deep learning. We introduce AraQReCC, a large-scale Arabic CQA dataset containing 9K conversations with 62K question-answer pairs, created by translating a subset of the QReCC dataset. To ensure data quality, we used COMET-based filtering and manual ratings from large language models (LLMs), such as GPT-4 and LLaMA, selecting conversations with COMET scores, along with LLM ratings of 4 or more. AraQReCC facilitates advanced research in Arabic CQA, improving clarity and relevance through question rewriting. We applied AraT5 for question rewriting and used BM25 and Dense Passage Retrieval (DPR) for passage retrieval. AraT5 is also used for question answering, completing the end-to-end system. Our experiments show that the best performance is achieved with DPR, attaining an F1 score of 21.51% on the test set. While this falls short of the human upper bound of 40.22%, it underscores the importance of question rewriting and quality-controlled data in enhancing system performance.

1 Introduction

Conversational Question Answering (CQA) enables systems to provide contextually relevant answers across multi-turn dialogues, with applications in virtual assistants, customer support, and information retrieval (Reddy et al., 2019). Unlike single-turn QA, CQA systems must maintain conversational context and handle implicit references to previous exchanges.

While substantial research exists for English CQA (Reddy et al., 2019; Qu et al., 2020; Anantha et al., 2021; Choi et al., 2018), Arabic one of the world's most widely spoken languages lacks effective CQA systems. This gap stems from Arabic's linguistic complexity and the absence of high-quality datasets, limiting accessibility for Arabic speakers.

We address this gap by introducing the first open-domain Arabic CQA system with question rewriting. Our approach leverages translated datasets with rigorous quality control to tackle Arabic-specific challenges.

To achieve this, we created AraQReCC, a large-scale Arabic CQA dataset, by translating a subset of the English QReCC dataset (Anantha et al., 2021). AraQReCC contains 9K conversations and 62K question-answer pairs. The QReCC dataset is chosen based on its proven effectiveness in question rewriting (Vakulenko et al., 2021), a crucial component for conversational QA.

For question answering and question rewriting, we use the AraT5 model (Elmadany et al., 2022), which has shown strong performance on Arabic NLP tasks. Additionally, we incorporate two retrieval methods BM25 and Dense Passage Retrieval (DPR) to retrieve relevant passages. Experiments on AraQReCC show similar trends to those observed in QReCC, highlighting the dataset's effectiveness.

To summarize, our contributions are:

- Creating the first Arabic conversational question answering dataset by translating the QReCC dataset with rigorous quality control measures. The created dataset is made publicly available to the research community.
- Applying comprehensive translation quality control using COMET-based filtering with balanced thresholds ($\geq$65% for training, $\geq$70% for development and test sets) and multiple large language models for rating, validated through human evaluation showing substantial agreement with GPT-4o ratings.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 84–96

Question	ماذا أدى إلى الجراحة؟ What led to the surgery?
Rewrite	ما الذي أدى إلى جراحة القلب المفتوح لنواز شريف؟ What led to Nawaz Sharif's open-heart surgery?
Answer	تدهور صحة نواز شريف أجبره على الخضوع لعملية قلب مفتوح قبل ثلاثة أيام فقط من تقديم الميزانية السنوية لباكستان. Nawaz Sharif's deteriorating health forced him to undergo an open heart surgery only three days before the presentation of Pakistan's annual budget.
Question	هل مات بنوبة قلبية؟ Did he die from a heart attack?
Rewrite	هل مات نواز شريف بنوبة قلبية؟ Did Nawaz Sharif die from a heart attack?
Answer	لا يزال نواز شريف على قيد الحياة ويقضي عقوبة بالسجن لمدة ١٠ سنوات منذ عام ٢٠١٨. Nawaz Sharif is still alive and serving a 10 year prison sentence since 2018.
Question	كيف كانت حياته العائلية؟ How was his family life?
Rewrite	كيف كانت حياة عائلة نواز شريف؟ How was Nawaz Sharif's family life?
Answer	تزوج نواز شريف من كلثوم نواز شريف وهي من أصل كشمير. Nawaz Sharif married Kalsoom Nawaz Sharif, who was also of Kashmiri descent.

Figure 1: Sample conversation from AraQReCC dataset.

- Developing an end-to-end system for open-domain Arabic CQA using established modules from prior work in open-domain QA and demonstrating the critical importance of question rewriting for system performance.

2 Background

Open-domain question answering (QA) systems aim to handle queries across diverse knowledge domains without being restricted to predefined topics. The introduction of conversational elements adds further complexity, as systems must maintain dialogue state and resolve contextual dependencies across multiple turns.

Conversational Question Answering (CQA) extends traditional QA by incorporating the dialogue context and previous interactions, enabling more accurate and contextually relevant responses. Unlike single-turn QA, CQA requires handling multi-turn conversations, where understanding user intent often involves resolving coreference, ellipsis, and pragmatic reasoning (Choi et al., 2018; Reddy et al., 2019). These challenges necessitate advanced techniques for dialogue modeling and context tracking.

In open-domain CQA, systems must interpret user queries within the evolving conversation, leveraging both prior dialogue history and large-scale knowledge sources. This involves retrieving relevant passages, reasoning over them, and generating contextually appropriate answers (Ma et al., 2023). The task has gained significant attention due to its applications in virtual assistants, customer support, and conversational AI platforms, where natural and interactive communication is essential.

Our work focuses on building an end-to-end system for open-domain CQA in Arabic. To this end, we translate an English dataset and adapt state-of-the-art methods originally developed for English (Qu et al., 2020). By leveraging these approaches, we aim to enable natural language interactions and

support knowledge dissemination in Arabic.

3 Related Work

Question answering research has progressed from single-turn open-domain QA to conversational settings that require maintaining context and resolving ambiguities. Recent work highlights three main directions: (i) open-domain QA methods for retrieval and comprehension, (ii) conversational QA approaches addressing coreference and ellipsis, and (iii) open-domain conversational QA, which combines large-scale retrieval with dialogue modeling and question rewriting. We review each of these directions below with emphasis on their relevance to Arabic QA.

3.1 Open-Domain Question Answering

Open-domain question answering refers to the task of automatically generating accurate and relevant answers to questions using a broad range of knowledge sources, without relying on specific predefined domains or contexts. Unlike open-domain conversational question answering it relies on one-turn questions (Reddy et al., 2019), (Choi et al., 2018), (Abdallah et al., 2024), (Yassine and Gammoudi, 2025), (Atef et al., 2020).

Several approaches address single-turn open-domain Arabic QA. Mozannar et al. (Mozannar et al., 2019) created the Arabic Reading Comprehension Dataset (ARCD) with 1,395 questions from Wikipedia articles. Their SOQAL system employs hierarchical TF-IDF retrieval and BERT-based reading comprehension (Devlin et al., 2018), achieving F1 scores of 61.3 for the reader and 27.6 for the complete system.

Almiman et al. (Almiman et al., 2020) proposed a deep neural network ensemble for Arabic CQA answer ranking, integrating lexical, semantic, and BERT-based features. Alsubhi et al. (Alsubhi et al., 2022) incorporated Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) to retrieve relevant passages from Wikipedia, using AraELECTRA (Antoun et al., 2020) for answer extraction. Their DPR approach outperformed traditional Arabic QA methods on both ARCD (Mozannar et al., 2019) and TyDiQA-GoldP (Clark et al., 2020) benchmarks.

3.2 Conversational Question Answering

Several English datasets have enabled progress in CQA, such as CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018). CoQA dataset is a valuable asset for constructing Conversational Question Answering systems. It consists of 127k conversational questions and their respective answers, collected from 8k conversations covering a wide range of domains. QuAC is an extensive dataset that focuses on Question Answering in Context. It consists of 14K dialogs where information-seeking questions are asked, resulting in a total of 100K questions.

There are several approaches for the CQA task. The first is by using full conversation history where the model incorporates inter-attention and self-attention mechanisms to comprehend the context and extract relevant information from the passage (Zhu et al., 2018). The second is by selecting history turns (Qu et al., 2019). The authors propose a method called history answer embedding to effectively incorporate conversation history into Conversational Question Answering (ConvQA) models. This approach simplifies the modeling of conversation history while achieving significant improvements in ConvQA. The third is by using question rewriting (Ye et al., 2023; Sekulic et al., 2024; Ye et al., 2024; Chen et al., 2022; Iovine et al., 2022) which aims to transform ambiguous questions into unambiguous ones, regardless of the surrounding conversation context (Vakulenko et al., 2021).

Question rewriting is a subtask that is trained separately, by taking the previous conversation history and rewriting the question accordingly. The Top two datasets for this task are CANARD (Elgohary et al., 2019) and QReCC (Anantha et al., 2021) datasets. CANARD dataset consists of 40K questions derived from the QuAC dataset. QReCC dataset includes rewritten versions of the entire QuAC dataset, in addition to extra data from other datasets.

3.3 Open-Domain Conversational Question Answering

Although there is a lack of research in Arabic conversational question answering, there is a lot of work in English language. Previous research in open-domain conversational question answering (CQA) for English has relied on repurposing existing datasets from the field of CQA.

The OR-QuAC dataset (Qu et al., 2020) is generated from QuAC and CANARD by replacing the original first question in QuAC (Choi et al., 2018) with the re-written question obtained from CANARD (Elgohary et al., 2019). For an open-

retrieval setting, they created a collection of over 11M passages using the whole Wikipedia corpus. The authors used the dataset to build an end-to-end system that incorporates a retriever, reranker, and reader based on Transformers. They demonstrate the significance of a learnable retriever and the benefits of history modeling across system components.

The QReCC dataset (Anantha et al., 2021) is a comprehensive open-domain CQA and question rewriting dataset that comprises conversations from various sources, including QuAC (Choi et al., 2018), TREC CAsT (Dalton et al., 2020), and Natural Questions (NQ) (Kwiatkowski et al., 2019). They created a collection of 10M web pages split into 54M passages. The authors extend BERTserini (Yang et al., 2019), an efficient method for open-domain question answering, by incorporating a question rewriting model that integrates conversational context.

Set Split	Train Set	Dev Set	Test Set	Overall
Full Dataset	40,221	10,139	12,389	62,749
COMET	7,537	1,782	2,190	11,509
LLM Rating	31,457	7,701	9,483	48,641
Dual Quality	6,341	1,500	1,850	9,691

Table 1: Number of Turns for Different Splits of AraQReCC Dataset

4 Dataset Creation

To simplify document collection, we translated conversations from the QuAC dataset (Choi et al., 2018), which draws primarily from Wikipedia and constitutes most of the QReCC dataset (Anantha et al., 2021). Using the Googletrans API[1], we created a dataset of 9K conversations with 62K question-answer pairs, split into training, development, and test sets.

We applied two quality control approaches to ensure translation quality:

- **COMET-based Filtering**: In the first approach, we used COMET (Crosslingual Optimized Metric for Evaluation of Translation) (Rei et al., 2020) to evaluate translation quality for each conversation. COMET is a neural machine translation evaluation metric that correlates well with human judgments and provides more nuanced assessment than traditional metrics like BLEU or ROUGE. To maintain a balanced dataset across splits, we applied different thresholds: conversations with COMET scores $\geq 65\%$ were selected for the training set, while conversations with COMET scores $\geq 70\%$ were selected for development and test sets. This approach ensures high-quality translations while maintaining sufficient training data volume.
- **LLM Rating**: In the second approach, we used large language models (LLMs) to evaluate the quality of the translation (Feng et al., 2021). Specifically, we employed GPT-4o, LLaMA 3.1 70B, and LLaMA 3.1 405B to rate each translated conversation on a scale from 0 to 5. We then took the average score of all the models, and conversations with an average rating of 4 or higher were selected.
- **Dual Quality of COMET and LLM Rating**: Finally, we created a dataset split by taking the intersection of the conversations that passed both the COMET threshold and the LLM rating threshold (COMET $\geq 65\%$ for training, $\geq 70\%$ for dev/test, and LLM Rating ≥ 4).

To evaluate the consistency of the ratings provided by the LLMs, we computed Cohen's Kappa scores for the pairwise agreements between the models. The Kappa score between GPT-4o and LLaMA-3.1-70b is 0.25, indicating fair agreement, while the score between GPT-4o and LLaMA-3.1-405b is 0.38, reflecting moderate agreement. Additionally, LLaMA-3.1-70b and LLaMA-3.1-405b demonstrated a Kappa score of 0.49, also suggesting moderate agreement. These scores highlight a fair to moderate level of consistency, particularly between the two LLaMA models, suggesting reasonable reliability in the ratings. By leveraging multiple models for the rating process, we aimed to minimize subjectivity and provide a more robust evaluation of the translation quality.

To further validate our quality control approach, we conducted human evaluation on 1200 randomly sampled conversations from the test set. The evaluation was carried out by independent annotators who are native Arabic speakers with advanced proficiency in English, ensuring reliable assessment across both languages. Annotators rated translation quality using the same 0–5 scale employed by the LLMs. The distribution of human ratings is as follows: 0 ratings (0 samples), 1 rating (10 samples), 2 ratings (68 samples), 3 ratings (216 samples), 4 ratings (370 samples), and 5 ratings

[1] https://pypi.org/project/googletrans/

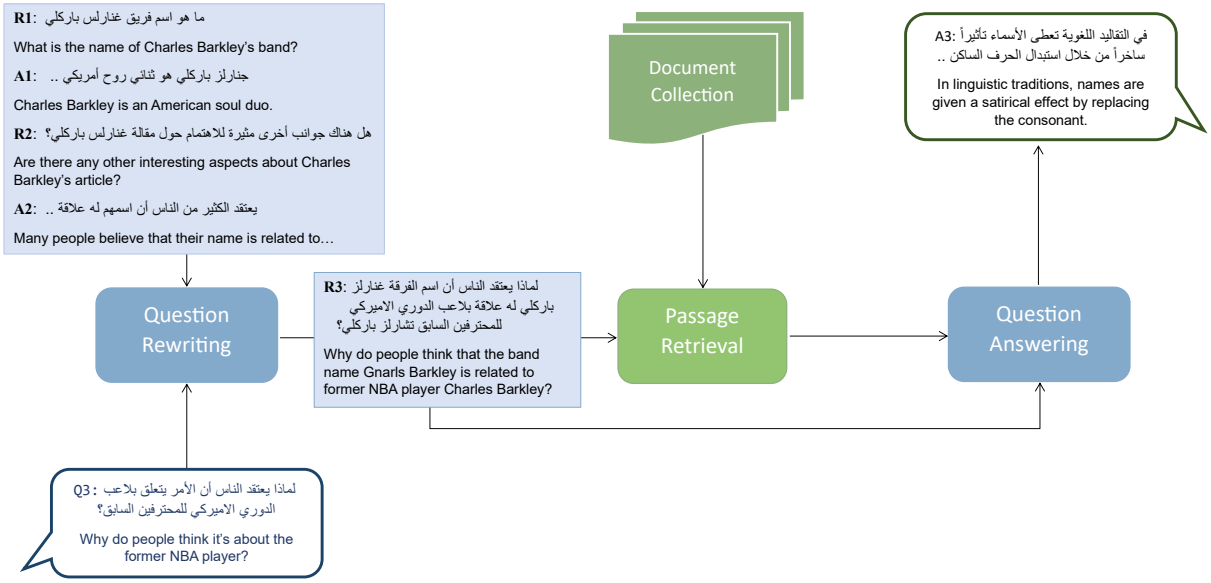

Figure 2: Overview of our end-to-end open-domain conversational question answering system. The pipeline begins with a user query (Q_3), which is rewritten into a contextually complete form (R_3) using the dialogue history. The rewritten query is then passed to the passage retrieval module (BM25 or DPR) to identify relevant passages, and finally to the answer generation module, which produces the response (A_3). This process ensures that ambiguous or context-dependent questions are clarified before retrieval, improving overall accuracy.

(536 samples), showing that the majority of translations (75.5%) received ratings of 4 or higher from human evaluation.

We computed Cohen's Kappa scores to measure agreement between human ratings and each LLM: GPT-4o achieved $\kappa = 0.725$ (substantial agreement), LLaMA-3.1-405b achieved $\kappa = 0.350$ (fair agreement), and LLaMA-3.1-70b achieved $\kappa = 0.263$ (fair agreement). These results demonstrate that GPT-4o shows the strongest correlation with human judgment, while the LLaMA models exhibit more moderate agreement. This validation confirms the reliability of our LLM-based quality assessment approach, particularly the effectiveness of GPT-4o ratings in identifying high-quality translations.

Table 1 provides the breakdown of the number of turns for the different splits of the AraQReCC dataset, including the full dataset, COMET split, LLMs rating split, and the dual quality split.

5 Document Collection

We use the entire Arabic Wikipedia corpus to construct a document collection since the passages in QuAC (Choi et al., 2018) are from Wikipedia. We extract the textual content from the wiki pages and split the texts into passages containing at least 220 tokens. We use the Arabic Wikipedia dump from 6/4/2023. As not all English Wikipedia pages are available in Arabic, we translate the English Wikipedia passages in QuAC to Arabic and add them to our collection. Finally, we end up with a collection of 9M passages. To assess translation quality, we manually reviewed a random sample of 100 translated passages, achieving an average human rating of 4.2/5.0 with 89% of passages rated 4 or higher for semantic accuracy and fluency.

6 Approach

Our end-to-end open-domain question answering system is illustrated in Figure 2. Given a user's original query Q_3, the system first rewrites it into a self-contained version R_3 that incorporates the necessary conversational context. This rewritten query is then used for passage retrieval and answer generation, producing the final answer A_3. By clarifying underspecified questions through rewriting, the system improves retrieval accuracy and ensures more relevant responses.

The rewritten question is then passed to the passage retrieval module, which searches a large document collection for relevant information. We employ retrieval models that encode queries and documents into a shared vector space for efficient similarity matching. The retrieved passages are then processed by the answer generation module, which

Model	Metric	Full Dataset		COMET		LLM Rating		Dual Quality	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
AraT5 (Full Dataset)	ROUGE1-R	65.45	64.86	70.08	69.57	67.50	67.07	71.34	71.36
	ROUGE1-P	**75.12**	**74.97**	**75.92**	**76.02**	**76.59**	**76.47**	**76.94**	**77.60**
	ROUGE1-F1	68.38	67.99	71.77	71.46	70.38	70.11	72.89	73.19
AraT5 (COMET)	ROUGE1-R	65.23	64.77	72.46	71.59	67.48	67.04	73.77	73.51
	ROUGE1-P	72.51	72.13	75.66	75.21	74.67	74.32	76.71	76.91
	ROUGE1-F1	67.40	67.00	73.07	72.31	69.61	69.24	74.27	74.17
AraT5 (Dual Quality)	ROUGE1-R	63.74	63.23	71.01	70.23	66.16	65.61	72.43	72.17
	ROUGE1-P	72.20	71.76	75.83	74.78	74.54	74.12	76.89	76.67
	ROUGE1-F1	66.42	65.97	72.35	71.44	68.82	68.35	73.62	73.39
AraT5 (LLM Rating)	ROUGE1-R	**69.01**	**68.94**	**74.37**	**73.84**	**71.22**	**70.89**	**75.44**	**75.54**
	ROUGE1-P	74.18	74.30	75.57	75.55	75.96	76.06	76.31	77.03
	ROUGE1-F1	**70.26**	**70.29**	**73.98**	**73.69**	**72.29**	**72.18**	**74.89**	**75.30**

Table 2: Question rewriting ROUGE1 scores (%) on development and test sets.

produces a concise and accurate response. Depending on the model, answers are either extracted directly from the retrieved text or generated in natural language. By integrating these components, our system enhances retrieval accuracy and ensures contextually relevant answers in an open-domain setting.

6.1 Question Rewriting

We use AraT5-base model (Elmadany et al., 2022) for question rewriting. To fine-tune it, we employ the history context from AraQReCC, which consists of the human-rewritten questions with the corresponding answers. The history context with the original question serves as the model input, while the rewritten question acts as the model output during the fine-tuning process. The hyperparameters we employ include 50 epochs, a batch size of 16, a learning rate of 3e-5, a maximum input length of 512, and a maximum target length of 128. The final model is selected based on the model checkpoint that achieved the highest ROUGE1 score on the development set.

6.2 Passage Retrieval

In our study, we incorporate two retrieval models: BM25 (Robertson et al., 1995) and the DPR retriever (Karpukhin et al., 2020). BM25 employs a bag-of-words scoring function to rank documents for a given query. In contrast, DPR Retriever learns dense vector representations of documents and queries, utilizing the dot product between them as a ranking function.

We use Anserini (Yang et al., 2017) for indexing BM25. After experimenting with various parameters for BM25, we found that the best results were achieved using the BM25 model with $k_1 = 0.9$ and $b = 0.4$.

To train the DPR (Dense Passage Retrieval) model, we construct a dataset by utilizing the QuAC passages as positive context. Additionally, we incorporate the top passages retrieved from BM25 with a top-30 selection as negative context. In (Alsubhi et al., 2022), the authors have demonstrated that fine-tuning mDPR on Arabic datasets produces promising results. Therefore, we fine-tuned our DPR model on the filtered dataset using the weights of a Multilingual DPR Model based on bert-base-multilingual-cased (Devlin et al., 2018) from huggingface[2], leveraging the Haystack library[3]. When fine-tuning our DPR model, we utilize the following hyperparameters: a maximum query length of 64, a maximum passage length of 512, 4 epochs, a batch size of 12, and 2 gradient accumulation steps. The final model is selected based on the model checkpoint that achieved the highest F1 score on the development set. Then we use our fine-tuned passage encoder to encode our passages collection and index them using FAISS flat index (Johnson et al., 2019).

6.3 Question Answering

We use AraQReCC dataset to fine-tune AraT5-base model for question answering. We use rewritten

[2] https://huggingface.co/voidful
[3] https://haystack.deepset.ai/

Model	Question	Full Dataset		LLM Rating		COMET		Dual Quality	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
BM25	Original Question	5.01	5.36	5.13	5.80	2.23	2.52	2.37	2.77
	Rewrite Full Dataset	28.09	28.63	30.20	30.99	18.49	17.52	20.03	14.42
	Rewrite LLM Rating	32.02	32.84	34.33	35.46	20.98	19.26	22.44	15.95
	Rewrite COMET	28.42	28.99	30.58	31.40	19.41	18.45	20.74	15.03
	Rewrite Dual Quality	26.77	26.94	29.16	29.62	17.78	17.08	19.17	14.36
	Gold Rewrite	38.88	40.18	41.43	42.49	24.15	23.83	25.27	<u>26.05</u>
DPR	Original Question	6.14	6.13	6.55	6.80	3.81	3.78	4.11	4.22
	Rewrite Full	<u>42.11</u>	<u>41.13</u>	42.23	41.03	28.22	25.63	31.07	19.93
	Rewrite LLM Rating	40.18	39.78	<u>44.52</u>	<u>43.87</u>	<u>29.50</u>	<u>26.59</u>	<u>32.51</u>	21.06
	Rewrite COMET	37.78	37.17	41.85	41.25	28.17	25.94	30.94	20.25
	Rewrite Dual Quality	37.12	35.89	41.43	39.98	27.50	25.00	30.40	19.29
	Gold Rewrite	**47.03**	**46.20**	**51.94**	**50.85**	**35.61**	**33.35**	**38.49**	**36.64**

Table 3: Mean reciprocal rank (MRR) scores (%) on development and test sets for Top-100 retrieval. Gold Rewrite refers to human-written reference rewrites that serve as the upper bound for question rewriting performance. The best scores are in bold, and the second-best scores are underlined.

questions along with their corresponding passages as the model input, and the model generates answers as the output. The hyperparameters we employ include 40 epochs, a batch size of 16, a learning rate of 5e-5, a maximum input length of 512, and a maximum target length of 128. The final model is selected based on the model checkpoint that achieved the highest F1 score on the development set.

7 Results and Discussion

Dataset Size Effects on Question Rewriting. As shown in Table 2, models trained on the full dataset achieve strong F1 scores (68.38% dev, 67.99% test), demonstrating the value of large-scale training data. However, the LLM-rated subset slightly outperforms the full dataset (70.26% dev, 70.29% test), suggesting that data quality can compensate for reduced quantity.

The COMET-filtered dataset achieves competitive results with balanced thresholds ($\geq$ 65% training, $\geq$ 70% dev/test). This approach maintains quality while preserving sufficient training volume. The dual quality split, combining COMET scores and LLM ratings, yields strong results by leveraging both automatic metrics and human-like assessment. Human evaluation validates this approach, showing substantial agreement with GPT-4o ratings ($\kappa = 0.725$).

Question Rewriting Impact on Performance. Question rewriting significantly improves both BM25 and DPR retrieval performance (Table 3). For example, BM25 improves from 5.01% to 32.02% MRR using LLM-rated rewrites, while DPR achieves 44.52% MRR, consistently outperforming BM25 across all splits. Gold rewrites establish upper bounds of 46.20% (DPR) and 40.18% (BM25).

For end-to-end QA (Table 4), the Gold Passage + AraT5 configuration performs best, reaching 21.51% F1 with LLM-rated rewrites and 23.85% F1 with gold rewrites. While substantial, these results fall short of the 40.22% human upper bound, highlighting remaining challenges in Arabic conversational QA. DPR consistently outperforms BM25, and question rewriting proves essential across all configurations.

8 Conclusion

In this work, we introduced AraQReCC, the first open-domain Arabic conversational question answering dataset, and demonstrated the importance of both data quality and question rewriting for enhancing retrieval and question-answering performance. Our quality control methodology, validated through human evaluation with substantial agreement between GPT-4o and human ratings ($\kappa = 0.725$), provides a reliable framework for

Model	Question	Full Dataset		LLM Rating		COMET		Dual Quality	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
BM25 + AraT5	Original Question	8.88	8.73	8.96	8.86	10.72	10.68	10.55	10.55
	Rewrite Full Dataset	17.62	16.99	17.16	17.32	17.62	16.99	17.62	12.57
	Rewrite LLM Rating	17.16	17.32	17.94	18.09	18.20	17.40	18.29	13.02
	Rewrite COMET	13.98	13.93	14.65	14.48	17.43	17.28	17.39	12.68
	Rewrite Dual Quality	13.80	13.80	14.54	14.46	17.03	17.22	17.18	12.43
	Gold Rewrite	16.19	16.22	17.14	17.05	19.73	20.45	19.65	20.85
DPR + AraT5	Original Question	10.34	10.34	10.60	10.63	12.48	12.18	12.35	12.32
	Rewrite Full Dataset	18.74	18.72	18.81	18.52	18.50	18.72	18.74	13.69
	Rewrite LLM Rating	18.41	18.67	18.81	18.52	18.96	18.65	19.40	13.71
	Rewrite COMET	14.68	14.89	15.49	15.64	17.90	18.83	18.17	13.63
	Rewrite Dual Quality	14.62	14.62	15.54	15.76	18.33	18.31	18.49	13.34
	Gold Rewrite	16.42	16.26	17.43	17.08	20.80	22.19	21.22	22.72
Gold Passage + AraT5	Original Question	20.65	19.56	20.43	20.03	12.05	11.86	17.61	17.67
	Rewrite Full Dataset	<u>22.30</u>	21.27	<u>22.67</u>	21.81	<u>22.93</u>	<u>21.73</u>	<u>23.01</u>	<u>21.86</u>
	Rewrite LLM Rating	22.07	<u>21.51</u>	22.63	<u>22.08</u>	22.67	21.22	22.36	22.36
	Rewrite COMET	10.61	10.35	10.88	10.67	13.59	12.93	13.71	13.06
	Rewrite Dual Quality	16.32	15.92	15.69	15.29	20.30	18.83	20.47	19.01
	Gold Rewrite	**25.35**	**23.85**	**25.89**	**24.29**	**24.80**	**24.69**	**24.89**	**24.93**
Extractive Upper Bound		40.31	40.22	39.84	39.76	39.46	38.58	39.73	38.79

Table 4: Question answering F1 scores (%) across different models and dataset splits for development and test sets. Bold values indicate the best scores, while underlined values represent the second-best scores.

assessing translation quality in low-resource languages.

The results of our experiments revealed that question rewriting plays a critical role in boosting the performance of both BM25 and DPR retrieval models. DPR consistently outperforms BM25 across all dataset splits, with the best F1 scores achieved using LLM Rating-based rewrites and Gold Rewrites.

These findings underscore the importance of both data quality control and question rewriting in open-domain conversational question answering systems. The combination of high-quality rewrites and optimized retrieval models is key to achieving better performance. Future work should focus on further optimizing passage retrieval and refining question rewriting techniques to close the gap between automated systems and human-level performance. Measuring performance against state-of-the-art large language models will also be considered for future work. We will release AraQReCC publicly to encourage further research on Arabic conversational QA.

Limitations

One notable limitation of our approach is the use of translated data. While the AraQReCC dataset provides a valuable resource for the Modern Standard Arabic conversational question answering, it may not capture the nuances and variations present in different Arabic dialects. As a result, the performance of our system on Arabic dialects might be suboptimal. Future work should aim to incorporate more diverse and region-specific data to improve the system's performance on Arabic dialects.

Overall, while our system shows promising results for open-domain Arabic conversational question answering, it faces some challenges in accurately retrieving and generating answers, particularly when confronted with ambiguous questions.

References

Ahmed Abdallah, Mahmoud El-Haj, and Mohammad Al-Tawfiq. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. *arXiv preprint arXiv:2403.01234*.

Abeer Almiman, Ola Abutayeh, and Fadi Al-Hussaini. 2020. Deep neural network approach for arabic community question answering. In *Proceedings of the 2nd International Conference on Advanced Machine Learning Technologies and Applications (AMLTA 2020)*, pages 1–9. Springer.

Kholoud Alsubhi, Amani Jamal, and Areej Alhothali. 2022. Deep learning-based approach for arabic open

domain question answering. *PeerJ Computer Science*, 8:e952.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

Adel Atef, Bassam Mattar, Sandra Sherif, Eman Elrefai, and Marwan Torki. 2020. Aqad: 17,000+ arabic questions for machine comprehension of text. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Chen Chen, Wenliang Chen, Yang Li, Jiansheng Li, Jiaying Li, Peixin Shi, and Yanan Wang. 2022. Reinforced question rewriting for conversational question answering. *arXiv preprint arXiv:2208.01257*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1985–1988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. *Conference on Empirical Methods in Natural Language Processing*.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring dialogpt for dialogue summarization. *arXiv preprint arXiv:2105.12544*.

Stefano Iovine, Luca Fogliato, Matteo Gatta, Marco Giraudo, and Andrea Lanza. 2022. Cyclekqr: Unsupervised bidirectional keyword-question rewriting. *arXiv preprint arXiv:2209.07663*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *ACL 2019*, page 108.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of EMNLP*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Ivan Sekulic, Krisztian Balog, and Fabio Crestani. 2024. Towards self-contained answers: Entity-based answer rewriting in conversational search. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 209–218.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Marwa Yassine and Mounira Gammoudi. 2025. Eadbi-lstm-bert: a novel deep learning architecture for arabic question answering systems. *arXiv preprint arXiv:2501.01234*.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *arXiv preprint arXiv:2310.09716*.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2301–2305.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

A Answer Generation Results and Analysis

In this appendix, we provide additional insights into the question rewriting and the answer-generation process of our end-to-end system. We present tables showcasing the answers generated from the retrieved passages and analyze the system's performance.

Table 5, shows question rewriting model performance on a random sample from the test set. It compares the gold rewrite with the text generated by a question rewriting model. It presents several examples along with the ROUGE1-R scores without using stemmer, which in general, indicates the similarity between the generated rewrite and the gold rewrite.

The analysis reveals that the question rewriting model shows varying levels of performance in generating accurate rewritten questions in Arabic. While some generated rewritten questions closely match the gold rewritten questions and achieve high ROUGE1-R scores as in the first example with a score of 100%, others exhibit discrepancies and lower scores.

In some cases, the model partially captures the essence of the original question but introduces an incorrect reference as in the second example. In other cases, the model generates rewritten questions that capture the overall topic of the original text but include additional information as in the third example. Also, there are instances where the model falls short in reproducing all the specific details, such as names, associated with the given context as in the fourth example. Sometimes the model generates rewritten questions that diverge significantly from the gold rewritten questions and fail to convey the correct meaning as in the fifth example. Overall, the analysis shows that the question rewriting model's performance varies across different examples. While some generated texts closely match the original texts and achieve high scores, others exhibit discrepancies and lower scores, indicating the need for further improvements in capturing the intended meaning.

A.1 End-to-End System Analysis

Tables 6 and 7 demonstrate our system's performance with gold rewritten questions, revealing both capabilities and limitations.

Table 6 shows cases where both BM25 and DPR retrieve identical passages but generate inconsistent answers (Example 1: date discrepancies), and where both retrievers fail entirely (Example 2: wrong domain retrieval). These examples highlight challenges in accurate answer extraction and retrieval precision.

Table 7 reveals that multiple passages can contain correct answers for the same question. Notably, BM25 sometimes achieves higher F1 scores despite retrieving incorrect passages, suggesting that partially relevant documents can still provide useful information. This indicates the complexity of passage-answer relationships in Arabic conversational QA.

Rewrite	Text	ROUGE1-R %
Gold	هل كان هناك أي شيء آخر يمكنك مشاركته حول بيير مونتوكس؟ Was there anything else you could share about Pierre Monteux?	
Model	هل كان هناك أي شيء آخر يمكنك مشاركته حول مقال بيير مونتوكس؟ Was there anything else you could share about Pierre Monteux's article?	100
Gold	ما الذي يمكن أن تخبرني به أيضًا عن مهنة ستيفن روش المهنية في ركوب الدراجات؟ What else can you tell me about Stephen Roche's professional cycling career?	
Model	ما الذي يمكن أن تخبرني به أيضا عن مسيرة ستيفن روش المهنية إلى جانب عملية جراحية؟ What else can you tell me about Stephen Roach's career besides surgery?	66.67
Gold	عن ماذا كان الفيلم الوثائقي أمريكا هيلاري؟ What was the Hillary America documentary about?	
Model	ما هو الفيلم الوثائقي ، أمريكا هيلاري: التاريخ السري للحزب الديمقراطي؟ What is the documentary, Hillary's America: The Secret History of the Democratic Party?	42.86
Gold	كيف ماتت ستيلا دكوورث؟ How did Stella Duckworth die?	
Model	كيف مات ستيلا؟ How did Stella die?	33.33
Gold	هل كان المستقبل هو إعادة إطلاقه في العصور الوسطى؟ Was the future a re-launch of the Middle Ages?	
Model	هل أعيد إصدار الألبوم المستقبل؟ Will the future album be re-released?	11.11

Table 5: Question rewriting examples comparing model outputs with gold standard rewrites. ROUGE1-R scores measure semantic similarity between generated and reference rewrites, illustrating varying model performance across different contexts.

B Hyperparameter Tuning

We conducted grid search over key hyperparameters for AraT5 fine-tuning. We settled on the following values:

Question Rewriting: 50 epochs, batch size 16, learning rate 3×10^{-5}.

Question Answering: 40 epochs, batch size 16, learning rate 5×10^{-5}.

Early stopping was applied after 5 epochs without improvement and . Models were evaluated on development sets using ROUGE-1 for question rewriting and F1 for QA tasks. Final models were selected based on best development set performance.

C LLM Rating Prompt

To ensure consistent evaluation across GPT-4o, LLaMA-3.1-70b, and LLaMA-3.1-405b, we used a standardized prompt for translation quality assessment. The prompt requests numerical ratings (0-5 scale) without additional commentary to enable direct comparison.

The exact prompt used is:

```
Rate the following translation
on a scale from 0 (terrible)
to 5 (perfect), focusing on
these aspects: accuracy of
meaning, fluency and grammatical
correctness, proper handling of
names and terminology. Provide
only the numeric rating (0-5)
with no additional commentary.
Source: {source}
Translation: {translation}
Rating:
```

Gold Rewrite	ما هي المدة التي لعبها جيسون جيامبي في لعبة صكي سش؟ How long has Jason Giambi been with the Sky sox?
URL	https://en.wikipedia.org/wiki/Jason_Giambi
BM25 Predicted URL	https://en.wikipedia.org/wiki/Jason_Giambi
DPR Predicted URL	https://en.wikipedia.org/wiki/Jason_Giambi
Answer	لعب جيسون جيامبي في سوكس من ٢٣ أغسطس ٢٠٠٩ إلى ١ سبتمبر ٢٠٠٩ Jason Giambi played for the Sox from August 23, 2009 to September 1, 2009.
BM25 + AraT5 Answer	لعب جيسون جيامبي في صكي سش من ٢٠٠١ إلى ٢٠٠٣. Jason Gimby played for the Sky sox from 2001 to 2003.
DPR + AraT5 Answer	لعب جيسون جيامبي في صكي سش من ٢٠٠١ إلى ٢٠٠٣. Jason Gimby played for the Sky sox from 2001 to 2003.
BM25 + AraT5 Answer F1	52.17
DPR + AraT5 Answer F1	52.17
Gold Rewrite	ماذا كان رم؟ what was rem?
URL	https://en.wikipedia.org/wiki/Michael_Stipe
BM25 Predicted URL	https://ar.wikipedia.org/wiki/Sleep_medicine
DPR Predicted URL	https://en.wikipedia.org/wiki/LeAnn_Rimes
Answer	هي فرقة روك بديلة . It is an alternative rock band.
BM25 + AraT5 Answer	رم هو جهاز اختبار اختبار للقدرة العقلي على التحقق من صحة نتائج مذكرة النوم. rem is a test device for mental ability to validate sleep note results.
DPR + AraT5 Answer	رم هو الألبوم الخامس لمغنية موسيقى الريف الأمريكية ، وهو الألبوم الخامس لها. rem is the fifth studio album by the American country music singer.
BM25 + AraT5 Answer F1	0.00
DPR + AraT5 Answer F1	0.00

Table 6: End-to-end system performance with gold rewritten questions, showing cases where both retrievers find the same passage (top) or both fail to retrieve relevant passages (bottom), demonstrating some system limitations in answer consistency and retrieval accuracy.

Gold Rewrite	هل لعب إيان بوثام في سومرست؟ Did Ian Botham play for Somerset?
URL	https://en.wikipedia.org/wiki/Ian_Botham
BM25 Predicted URL	https://ar.wikipedia.org/wiki/Viv_Richards
DPR Predicted URL	https://en.wikipedia.org/wiki/Ian_Botham
Answer	لعب إيان بوثام معظم لعبة الكريكيت من الدرجة الأولى في سومرست. Ian Botham has played most of his first-class cricket for Somerset.
BM25 + AraT5 Answer	لعب إيان بوثام في سومرست من ١٩٩٢-١٩٨٣. Ian Botham played for Somerset from 1983-1992.
DPR + AraT5 Answer	لعب إيان بوثام في سومرست من ١٩٨٠ إلى ١٩٨٣. Ian Botham played for Somerset from 1980 to 1983.
BM25 + AraT5 Answer F1	66.66
DPR + AraT5 Answer F1	60.00
Gold Rewrite	هل حقق ألبوم وَزس الذي ألفه إنريكي إغليسياس أداءً جيدًا في الخارج؟ Did Enrique Iglesias' Quizás Album Do Well Abroad?
URL	https://en.wikipedia.org/wiki/QuizÃąs_(album)
BM25 Predicted URL	https://en.wikipedia.org/wiki/Triumph_(band)
DPR Predicted URL	https://en.wikipedia.org/wiki/Enrique_Iglesias
Answer	دخل الألبوم أيضًا في ٢٠٠ على قوائم الألبومات في المملكة المتحدة ، بالإضافة إلى الأداء الجيد عبر أمريكا اللاتينية حيث ذهب إلى البلاتين في مقاطعات مثل المكسيك والأرجنتين. The album also entered the top 200 on the UK album charts, in addition to performing well across Latin America where it went platinum in provinces such as Mexico and Argentina.
BM25 + AraT5 Answer	الألبوم ، تم ترشيح إنريكي إغليسياس لجائزة جرامي لأفضل ألبوم موسيقى الروك ، The album, Enrique Iglesias was nominated for a Grammy Award for Best Rock Album,
DPR + AraT5 Answer	باع الألبوم مليون نسخة في أسبوع ، مما جعلها الألبوم الأسرع مبيعا باللغة الإسبانية منذ The album sold one million copies in a week, making it the fastest-selling Spanish-language album since
BM25 + AraT5 Answer F1	5.12
DPR + AraT5 Answer F1	13.95

Table 7: Examples showing how different retrieval methods can find partially relevant passages. BM25 sometimes achieves higher F1 scores than DPR despite retrieving incorrect passages, indicating that multiple passages may contain relevant information for the same question.