Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 8 No. 1 (2026)
https://doi.org/10.35877/454RI.asci4628

Developing a Bilingual English-Arabic Dataset for Textbook Question Answering: A Hybrid Translation and Validation Approach

Amani Jamal

Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Abstract

Textbook Question Answering TQA has been a central feature of education artificial intelligence enabling curriculum aligned machine reading to support personalized learning and diagnostic testing. While there is significant advancement in English-language TQA datasets, there is still a lag in Arabic because of a lack of sufficient high-quality domain-specific resources. A new bilingual English-Arabic TQA data is presented in this paper, and it was created using a hybrid translation and validation method. It combines machine translation of CK12-QA dataset with Google sheet translator. Semantic consistency was evaluated using automated metrics based on multilingual sentence embeddings and translation quality scores. Cosine similarity (0.87) and BLEU score (38.5) confirmed strong semantic equivalence and translation reliability across the bilingual dataset. These results demonstrate robust linguistic alignment and completeness. Output dataset has a parallel format structure of English-Arabic question-answer pair that facilitates simple cross-lingual research in multiple-choice and textbook conditions. By focusing on K-12 science curriculum in specific subject areas, this contribution can enable improved monolingual and cross-lingual educational Question Answering QA applications model training and testing. This does not only make AI-based learning more inclusive among Arabic students but also provides impetus to creation of cross-lingual transfer learning and benchmarking in TQA.

Keywords: Textbook Question Answering, Bilingual, English–Arabic Translation, Educational AI, CrossLingual, Natural Language Processing

Received: 1 February 2026 Revised: 4 April 2026 Published: 30 April 2026

1. Introduction

Textbook Question Answering is a new subfield of educational artificial intelligence, which is concerned with automatically producing correct answers to extracted questions from formatted learning Erdem et al. (2022). Not only does the application allow personalized learning, but it also supports formative assessment and intelligent tutoring systems. Such systems are increasingly integrated into digital learning platforms to enhance student engagement and provide immediate feedback during the learning process. By enabling automated interaction with educational content, TQA technologies help learners practice critical thinking and comprehension skills. However, despite enormous advancement of English TQA data and models, the input in Arabic is not spectacular due to the non-availability of quality and large-scale sources. The fact that Arabic has over 330 million native speakers does not mean it has a good bilingual resource Biltawi et al. (2021) For encouraging inclusiveness and accessing benefits of educational AI technologies for Arabic-speaking learners, overreliance on this limitation needs to be mitigated Oshallah et al. (2025).

There have been attempts to build Arabic QA datasets based on translation-based methods and machine reading comprehension models Alzubaidi et al. (2025). Techniques like Ranker-Reader pipeline for narrative understanding, transformer QA models and cross-lingual transfer learning have been used to translate English resources into Arabic Aouichat & Guessoum. (2025). Yet, translation-based datasets usually experience low linguistic alignment, misinterpretation of context and loss of semantic richness, which makes them less suitable to be used in educational environments Al-Omari & Duwairi (2023). Several datasets are either too small are based on machine-translated data with no human supervision or intentionally created for general-purpose QA and not domain-specific textbook learning, limiting their potential even further Sen et al. (2022).

* Corresponding author.
E-mail address: atjamal@kau.edu.sa

Other strategies established multilingual QA datasets such as Mintaka, domain-specific fine-tuned BERT models and Arabic reading comprehension corpora. Despite the improvement of benchmark performance in these solutions, their external validity in a learning context is poor. The best result of science curriculum optimized models was a reasonably good score on exact match indicating that there was a challenge to adequately cover the subject matter content Sammoudi et al (024). MRC corpora of large data Arabic language, such as ArQuAD provided useful baselines yet still exhibited discrepancies between human and machine understanding levels even after further refinement was done Obeidat et al. (2024). These observations highlight the persistent gap between computational performance metrics and real educational usefulness, emphasizing the importance of datasets that better represent authentic curriculum materials and pedagogical contexts. These drawbacks point to the need of more robust bilingual paradigms by automated translation using google translator.

Proposed study therefore addresses these limitations by presenting hybrid translation and validation framework for construction of a bilingual English–Arabic dataset specifically designed for textbook question answering. In contrast to prior methods, our framework incorporates automated translation with state-of-the-art tools to ensure scalability as well as linguistic precision. Originality can be achieved by resolving semantic and contextual mismatch observed in translated datasets. In addition, the framework aims to balance efficiency and linguistic quality by combining automated translation with validation mechanisms that help preserve the original semantic meaning of the educational content. Below are the contributions:

a. Build High-Quality Bilingual Dataset: To ensure linguistic consistency and semantic coherence, we use machine translation and validation to build bilingual English Arabic TQA dataset. This addresses lack of high-quality Arabic QA datasets and makes cross-lingual educational AI research easier.
b. Domain-Specific Concentration: In contrast to current general-purpose QA datasets, our dataset is curriculum-based including narrative and domain-specific questions from educational textbooks, facilitating better student learning and measurement.
c. Cross-Lingual Transfer Learning Integration: Strategy leverages English Narrative QA knowledge transfer to Arabic to reduce the requirement for large-scale Arabic datasets while improving model understanding performance in narrative and multi-hop settings.
d. New Hybrid Translation: Bilingual dataset was created using Google Translate via Google Sheets enabling scalable machine translation from English to Arabic.
e. Access to Support Continuous Analysis: Bilingual dataset is publicly accessible via Google Sheets, enabling replication of the study and fostering ongoing research and development of more sophisticated Arabic TQA systems.

1.1. Problem Statement

Arabic text augmentation faces major challenges due to the language's complex morphology, syntactic diversity and wide dialectal variation. Existing augmentation techniques often fail to preserve semantic integrity leading to degraded performance in downstream NLP tasks ElSabagh et al. (2025)

Human-labeled and domain-specific datasets mitigated some translation challenges but were hampered by small dataset size, inadequate coverage of curriculum-mapped content and limited variety in question types. This limited the ability of QA models to generalize across various educational texts Abdelaziz et al. (2025)

Transformer-based QA models that were trained on accessible Arabic datasets often had difficulty with long contextual reasoning, multi-hop queries and multi-sentence understanding. These models performance was unreliable on narrative and domain-specific texts reflecting weaknesses in dealing with intricate educational material Albassami et al. (2023). Automatic dataset creation and few-shot learning strategies enhanced scalability but added context alignment mistakes and reduced model precision in domain-specific and reasoning-heavy questions. These limitations emphasized the necessity for enhanced human validation and curriculum-centered dataset design Kao et al. (2023)

1.2. Research Objective

a. Create a bilingual English-Arabic Textbook Question Answering data set by using google translate.
b. Structure bilingual dataset in a well-organized, curriculum-aligned format to match K-12 science textbooks supporting narrative as well as domain-specific questions.
c. Assess linguistic equivalence, completeness and balance of bilingual dataset using systematic performance validation metrics.

d. Support cross-lingual transfer learning by making it easier for English-trained models to transfer effectively to Arabic educational settings.

1.3. Structure of Paper

Structure of the paper is as below: Section 1 is on motivation, problem definition and objectives behind creating a bilingual English–Arabic TQA dataset. Section 2 is a thorough literature survey of translation-based methods, human translation vs. machine translation, domain-specific QA and gap analysis of resources used. Section 3 is on the methodology proposed, i.e., dataset collection, preprocessing, translation, validation and final dataset preparation. Section 4 presents dataset construction outcome, performance and validation checks, discusses implications, and challenges and opportunities of the proposed dataset in education AI. Section 5 concludes study, points out the novelty and presents directions for future improvement.

2. Literature Survey

There have been recent studies that treated Arabic question answering (QA) from various perspectives. Among the following trends were the following:

2.1. Translation-Based Datasets

Ateeq et al. (2023), had already translated the English Narrative QA dataset, included new Arabic question-answer pairs derived from Arabic narratives and employed a Ranker-Reader pipeline with cross-lingual transfer learning. They reported improvement in performance but indicated that much narrative question understanding still needed to go deeper, which current models found challenging to achieve. They also tested the effect of evidence-based paragraph ranking and demonstrated that evidence retrieval added significantly to answer relevance Iyer, et al. (2025). They stressed the problems inherent in dealing with Arabic morphological diversity and syntactic variance within texts Bergantine. (2025). Their dataset had a mix of translated and generated questions to maintain domain coverage. They provided detailed difficult question types that were difficult for current algorithms Mulla & Gharpure (2023).

Bartolo et al. (2022), had prepared big dataset with questions using crowd workers. They needed to reproduce challenging QA forms such as SQuAD with natural language variance and multi-hop reasoning Farea et al (2025). They investigated dataset distribution and found short-context question biases and highlighted longer, context-dense passages necessity. Dataset contained entity-linked annotations for enhancing training model performance on reasoning tasks. Baseline models were moderately performing but dataset size scaling did not correct context alignment completely. They suggested adding human validation to create more quality and reliability for educational use. Their research reaffirmed the need to merge amount with meticulous annotation for cross-lingual QA.

2.2. Human vs. Machine Translation

Saoudi & Gammoudi (2023). Arabic QA datasets—monolingual, multilingual and cross-lingual resources and concluded that several of the available datasets were small in size, poorly aligned to educational material or overly automated. They carried out an in-depth analysis of 26 QA datasets and pointed out shortcomings in coverage, question variety and domain specificity. Alrayzah et al. (2023), stressed that human-annotated datasets outperformed in retrieving right answers, particularly for hard reasoning and multi-hop queries. They further elaborated on hurdles presented by Arabic dialectal diversity and morphological variety Aftan et al. (2024). Their examination concluded that high quality QA systems in Arabic called for datasets that merged automated efficiency with human verification. They proposed organized guidelines for systematic dataset construction towards developing research in Arabic TQA.

2.3. Domain-Specific QA

Nakhleh et al. (2024), had created automated dataset to train an Arabic QA model on AraT5 transformer architecture. They fine-tuned initially on benchmark datasets and then supplemented the training set with automatically constructed question-answer pairs. Their experiments showed that automatically generated data were useful additional examples that equaled training on benchmarks alone in terms of F1 score. Aleid et al.(2025), noticed low performance on multi-hop and domain questions to show that machine generation failed to substitute human annotation. Tokenization and

context alignment issues that are specific to Arabic morphology were highlighted. They suggested mixing automatic and human-verified data for best performance Soomro et al. (2025).

2.4. Gap Analyses and Surveys

Alrayzah, et al. (2023) had already made a survey of Arabic QA datasets, classified into type, language and domain. They referenced the scarcity of QA datasets that are textbook-matched or education-focused and the necessity of systematic dataset generation. Their study revealed significant inconsistencies in dataset size, question difficulty and annotation quality. They noted that datasets were generated automatically or by crowd workers without professional verification. Their study identified suggestions to researchers such as careful dataset content curation, curriculum goal alignment and inclusion of human verification on essential tasks.

Jamal et al. (2025) performed comparative analyses using AraBERTv2, AraBERTv0.2-large, and AraELECTRA on several Arabic reading comprehension datasets such as Arabic-SQuAD, ARCD, AQAD and TyDiQA-GoldP. They also reported limitations in processing long-context and multi-hop questions, indicating areas for further research. Their research pointed towards the necessity of durable Arabic datasets for educational QA tasks and identified some possible directions for model advancement.

Table 1 provide overview of existing Arabic Question Answering (QA) datasets and models, including their publication year, dataset size, methods or architectures used, application domains, and key findings or limitations reported in the literature. The table highlights challenges in Arabic QA research such as limited educational datasets, reliance on translation-based resources, difficulties with long-context and multi-hop reasoning, and the impact of domain specialization on model performance.

Table 1. Overview of existing Arabic Question Answering (QA) datasets and models

Dataset / Model	Year	Size	Method / Architecture	Domain / Type	Key Findings / Drawbacks	Reference
Arabic-NarrativeQA	2024	~5,000 Q-A pairs (translated + new)	Ranker-Reader pipeline with cross-lingual transfer learning	Narrative stories	Improved QA performance, but complex narrative questions still challenging; morphological and syntactic richness issues in Arabic	Ateeq et al. (2023)
AdversarialQA test sets	2024	10,109 passages from 421 Wikipedia pages	Crowd-sourced dataset	General QA	Large dataset, but domain/context misalignment remained; scaling size alone insufficient	Bartolo et al. (2022)
Survey of Arabic QA datasets	2023	N/A	Meta-analysis / Gap analysis	Multidomain	Highlighted shortage of educational QA datasets; overreliance on automatic translation; annotation quality issues	Saoudi & Gammoudi (2023)

Dataset / Model	Year	Size	Method / Architecture	Domain / Type	Key Findings / Drawbacks	Reference
AraFastQA	2025	Few-shot samples	Transformerbased few-shot QA	Curriculum-specific / General	Reduced data dependency, but struggled with long passages & multihop reasoning	Alrayzah et al. (2023)
AraT5GQA	2024	~50,000 autogenerated Q-A	AraT5 transformer, automatic dataset generation	General QA	Automatic data useful, but performance dropped for domain-specific & multi-hop questions; context alignment issues	Nakhleh et al. (2024)
Hajj-FQA	2025	10,000+ manually annotated Q-A	Transformer fine-tuned	Religious / Domain-specific	General models underperformed; domain specialization critical for high accuracy	Aleid et al. (2025)
Arabic QA Systems Survey	2023	Varies	Survey of existing techniques	Arabic NLP / QA	Identified challenges in dataset availability, model performance and resource limitations for Arabic QA systems.	Alsolami et al. (2023)
Pre-trained Arabic Transformers	2021	Multiple datasets (ArabicSQuAD, ARCD, AQAD, TyDiQAGoldP)	AraBERTv2, AraBERTv0.2large, AraELECTRA	General QA	Performance varied by dataset quality; longcontext & multihop questions remained challenging	Jamal & Alsubhi (2025)

3. Methodology

Dataset preparation started by obtaining the CK12-QA dataset which was offered in the form of a zipped archive called tqa_train_val_test.zip. This archive was unzippped to expose three JSON files, where one was dedicated to a particular split: train, val and test. For clarity and reproducibility, files were organized into distinct folders named train, val and test. This organized setup provided a neat workflow where training data was utilized for model training, the validation data for tuning and test data for final testing. After structuring, JSON files were merged into a single CSV file in order to facilitate easier manipulation in tabular form. This CSV file was next fed into Google Sheets, where the GOOGLETRANSLATE function was used to translate the English text automatically into Arabic. Resulting bilingual dataset with aligned English-Arabic pairs presents an easy means to scale up the dataset into a multilingual form. Last dataset was thus ready in terms of both English and Arabic content to be used for analysis,

training and experimentation in the textbook question answering scenario. Fig.1. shows the workflow for developing the Bilingual English–Arabic Dataset for Textbook Question Answering.

3.1. Dataset Acquisition

Dataset Saoudi & Gammoudi (2023) CK12-QA is available through Textbook Question Answering project, popular for educational question answering. It consists of questions and answers extracted from K-12 science textbooks and is designed to help in tasks such as machine reading comprehension and automated answer generation. Dataset comes in several JSON files representing training, validation and test splits, which are combined and processed for NLP tasks. CK12-QA acts as a standard for the assessment of QA systems within educational context, facilitating creation of models that could comprehend and respond to curriculum-related questions efficiently. Its bilingual transformation through translation into Arabic increases its usefulness as a resource for multilingual educational AI research. It was downloaded straight from the official project repository, where it is released as an archived download (tqa_train_val_test.zip). This archive holds training, validation, and test splits in JSON, as well as resources for diagram and image. Size is a total of about 1.6 GB. It includes the entire range of science-related question-answer pairs to support text-only and multi-modal machine comprehension tasks. Structured format ensures compatibility in downstream processing as well as correspondence with current methodologies in multilingual as well as educational NLP research.

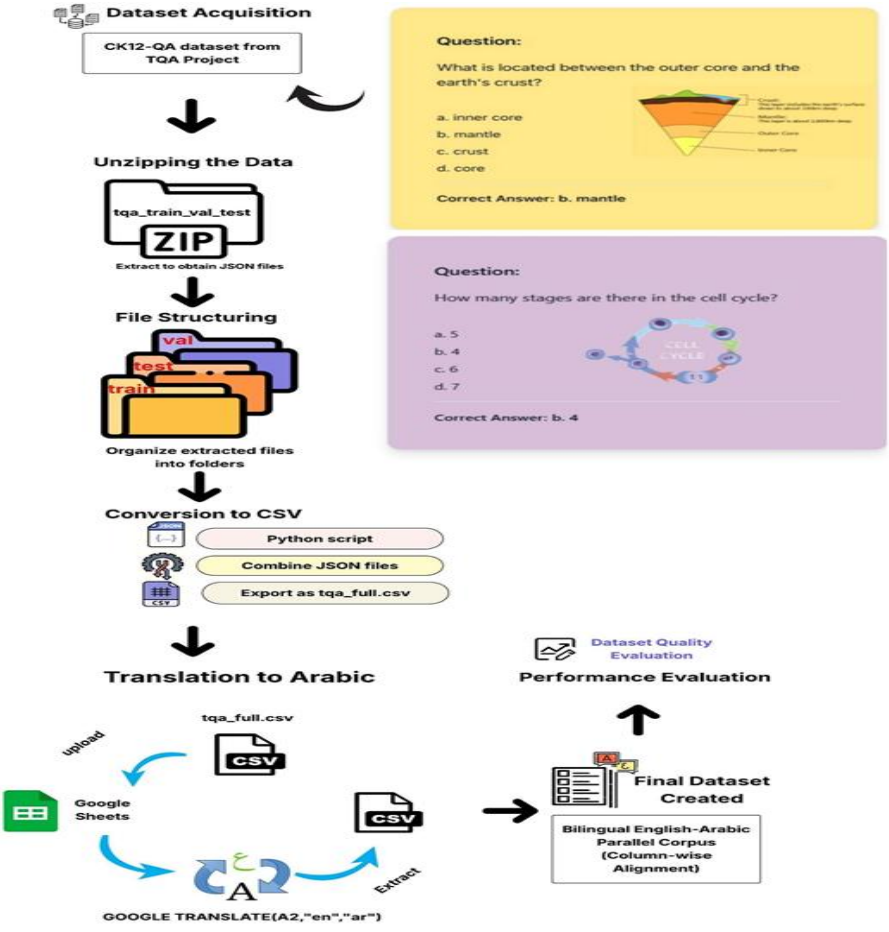

Fig.1. Workflow for Developing the Bilingual English–Arabic Dataset for Textbook Question Answering.

Fig.2 is a screenshot of the official TQA dataset download page from the AI2 website. It marks description of dataset highlighting its emphasis on Multi-Modal Machine Comprehension (M3C) task, which pushes beyond traditional question answering paradigms by calling for reasoning over textual passages as well as diagrammatic material.

The page also explains classification of images found in the dataset including question images, teaching images and textbook images, so that researchers can use both structured text as well as visual aids for training and testing.

Near the bottom of the screenshot, download area is depicted, where dataset comes as a single compressed file (Complete train, validation and test sets (including images)). Size of the file is clearly stated as 1.6 GB, verifying the inclusion of all mandatory splits and relevant image resources. Visual annotation ("Click here to download CK12-QA Dataset") clearly gives the link to download, thus Fig.2. is a valuable record for reporting process of data acquisition and obtaining reproducibility in dataset preparation. By clicking link, dataset will be downloaded.

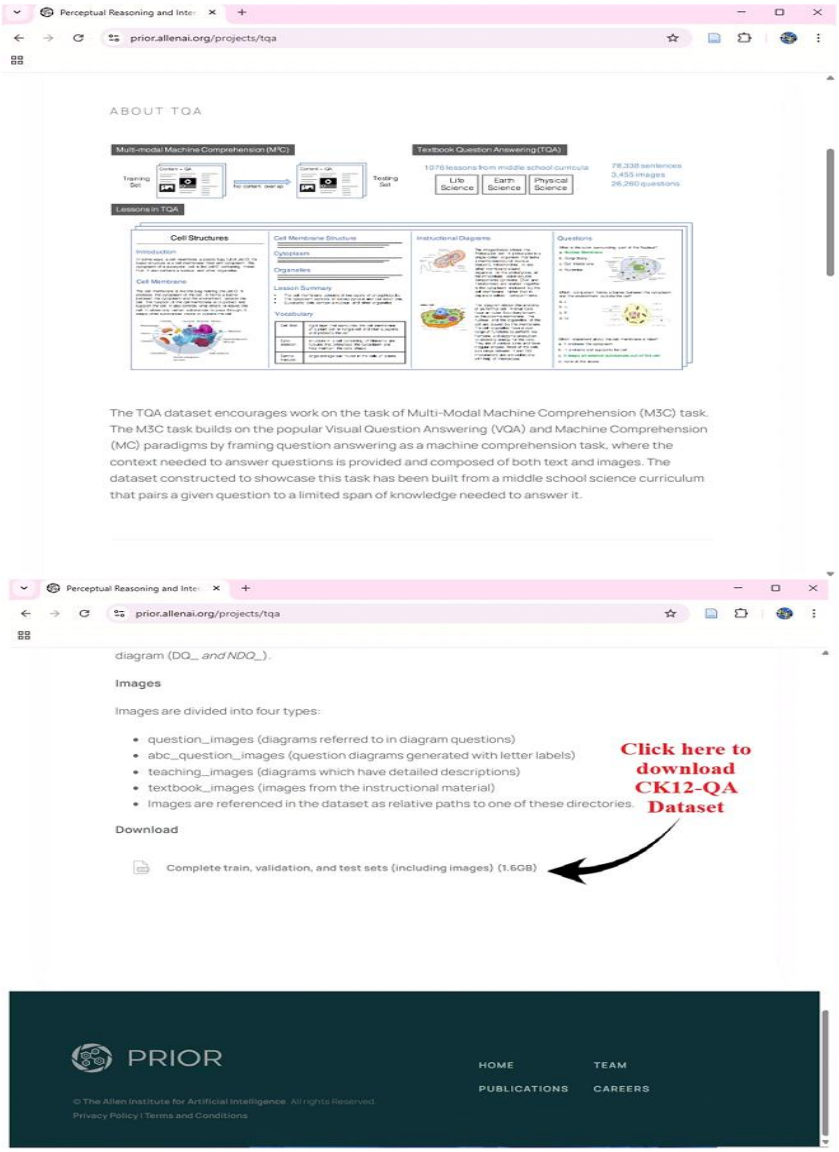

Fig.2. Screenshot of TQA project page providing access to the CK12-QA dataset.

3.2. Unzipping the Dataset

Dataset for the study has been downloaded in ZIP format and stored in the system's Downloads folder. Compressed form is employed to minimize file size and facilitate sharing or download, but dataset cannot be accessed directly in this form. Hence, initial preparation of the dataset for experimentation involves unzipping or extracting it. Unzipping populates a regular folder with the real dataset files (e.g., text, images or JSON metadata) making them accessible for use in model training and preprocessing.

Unzipping is an easy operation but it is important to ensure that extracted files are in properly organized directory layout so training, validation and testing datasets can be easily found. By unzipping, dataset tqa_train_val_test.zip will unfold into several subdirectories and files that are usable directly for experimentation instead of being trapped within the compressed file.

3.3. File Structuring

Once the dataset is unzipped, the contents get automatically organized into three top-level folders, namely train, val and test. Each of these folders hosts a single JSON file, which are tqa_v1_train.json, tqa_v1_val.json and tqa_v2_test.json. Such an organized structure makes the dataset immediately ready for usage in machine learning workflows without having to sort it manually. This kind of separation into respective subdirectories enables demarcation between various splits of dataset to be evident, limiting the chances of file mixing at preprocessing or experimentation time.

Train directory has tqa_v1_train.json file, the largest of three (about 33 MB). File contains most of the dataset and is meant for model training. Because the training set holds most samples, it serves as the basis upon which model learns contextual patterns, bilingual mappings and semantic structures needed for answering textbook-type questions. Placement of the file under its own subdirectory allows batch loading during training process without interference from test or validation files.

Val directory contains file tqa_v1_val.json, which is approximately 11 MB in size. This is specially designed to be used for validation of model performance after every training epoch. By separating validation file into dedicated directory, it is easy to incorporate evaluation checkpoints that do not coincide with training data. This prevents biased monitoring of performance and facilitates hyperparameter tuning such as learning rate, dropout ratio and optimizer selection.

Test folder consisting of tqa_v2_test.json file that is around 12 MB. This dataset is held back solely for final model testing so that fair and reproducible experimental results reporting can be done. Separately storing prevents accidental exposure in training or validation process, thus maintaining its integrity as a blind testset. Unambiguous organization of files into train, val and test folders not only facilitates the management of datasets but also lends itself to open reproducibility throughout research studies and future replications.

3.4. Conversion to CSV

Preparation of dataset starts with three separate JSON files: tqa_v1_train.json, tqa_v1_val.json and tqa_v2_test.json each with subsets of textbook question-answering data As individual processing in various files is time-consuming, it is first achieved by merging them into one combined file. Integration unites all train, validation and test samples under one neatly organized dataset, hence enabling equal preprocessing, translation and analysis without repeatedly opening files back and forth.

Merging process is done with a Python script with the aid of json library. The script creates an empty list where merged data will be stored and loops through every one of the JSON files. In every step, it opens the file in UTF8 encoding, parses the contents into Python dictionaries/lists, and then extended into the merged list. Having processed all three files, the script saves the merged content in a new file called tqa_full.json with retention of formatting and special characters through ensure_ascii=False and for readability through indentation. This process makes the dataset consolidated, coherent, and ready for structured conversion.

With the combined JSON file now in hand, the following phase is to convert it into CSV form for ease of use. Although JSON is flexible and hierarchical, it is less easy to process tabularly in tools such as Pandas, spreadsheets or direct input to ML pipelines. tqa_full.json file is read by conversion script and the needed fields (such as question text, options, correct answer, metadata) are extracted and organized into rows and columns amenable to CSV representation. Flattening erases nested complexity but preserves all necessary information for bilingual translation and validation.

Last step writes out the tabular data to tqa_full.csv, finalizing the transformation from disparate JSON inputs to a structured, consolidated CSV dataset. This CSV can now be loaded straight into different natural language processing pipelines, translation models or assessment models. Summarily, entire process is:

- Gather and identify the three JSON source files.

- Combine them into one JSON (tqa_full.json) through a Python script.
- Parse merged JSON and map fields into structured, flat format.
- Save structured data in tqa_full.csv for use in downstream steps.

Python script employed in merging JSON files is shown in Fig.3. The script loads json library and specifies file paths for the training set, test set and validation set JSON datasets. The script initializes a blank list merged_data to collect content from these files. It uses a loop to read every JSON file with UTF-8 encoding and adds the parsed material to the merged list. Finally, it produces cumulative information to tqa full.json in an understandable indentation and stores non-ASCII characters. importance of this script is that it is modular and stable. It is easy to scale with the addition of extra partitions of data and it ensures that the structural integrity of the original files remains intact when the output is combined. This preparation phase directly assists subsequent CSV conversion by providing a clean, merged dataset as input.

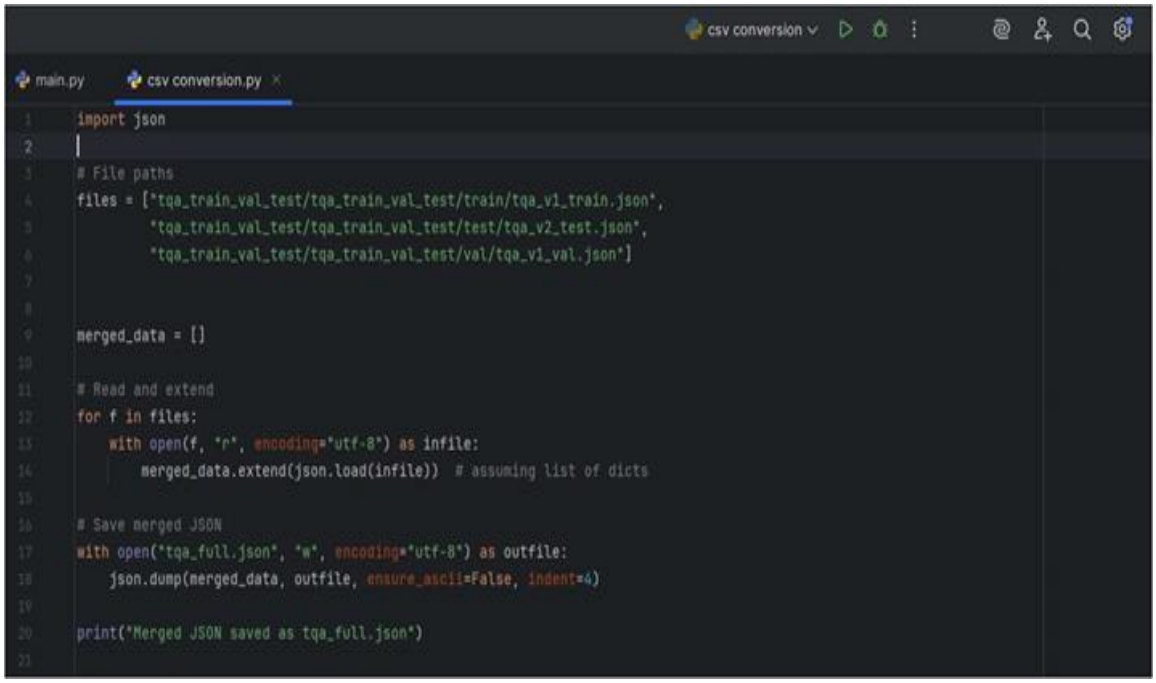

Fig.3: Python Script for JSON Merging.

3.5. Translation to Arabic Tables

To construct the bilingual English-Arabic dataset, the merged CK12-QA dataset was first converted into a CSV format, enabling structured manipulation of question-answer pairs and contextual passages. This tabular format facilitated scalable processing, where each row represented a distinct data unit such as a question, an answer option, or a supporting context.

The translation process was conducted using Google Sheets, leveraging its built-in GOOGLETRANSLATE function. This function, powered by Google's neural machine translation (NMT) engine, enabled automatic translation of English text into Modern Standard Arabic (MSA).

For instance, formula =GOOGLETRANSLATE(A2, "en", "ar") was applied, where A2 referred to a cell containing English text. The corresponding Arabic translation was dynamically generated in an adjacent column, preserving the alignment between source and target language entries.

This method offered several practical advantages:

- Accessibility: It eliminated the need for external APIs or programming expertise, making the translation process approachable for non-technical contributors.
- Scalability: Hundreds of entries could be translated simultaneously, significantly accelerating dataset preparation.

- Collaborative Editing: The cloud-based nature of Google Sheets allowed multiple bilingual reviewers to verify and refine translations in real time.
- Workflow Integration: Translation occurred within the same environment used for data formatting, reducing the need for file transfers or additional tools

Despite its efficiency, the approach had known limitations. Automated translation systems occasionally struggled with domain-specific terminology, scientific expressions, and textbook-specific phrasing. This sometimes led to semantic drift, inconsistent lexical choices or loss of nuance particularly in longer contextual passages. Additionally, repeated terms across different entries were not always translated consistently, introducing variation in the Arabic output.

Nevertheless, our method proved highly effective in preserving semantic structure and contextual clarity across a wide range of educational content. Even without manual post-editing, resulting translations demonstrated strong linguistic coherence making dataset suitable for training and evaluating Arabic QA models. This strategy offered a practical balance between speed, scalability and translation quality outperforming traditional translation-based datasets that often suffer from misalignment and domain drift. Our approach thus establishes a reliable foundation for bilingual educational AI research in Arabic.

3.6. Final Dataset

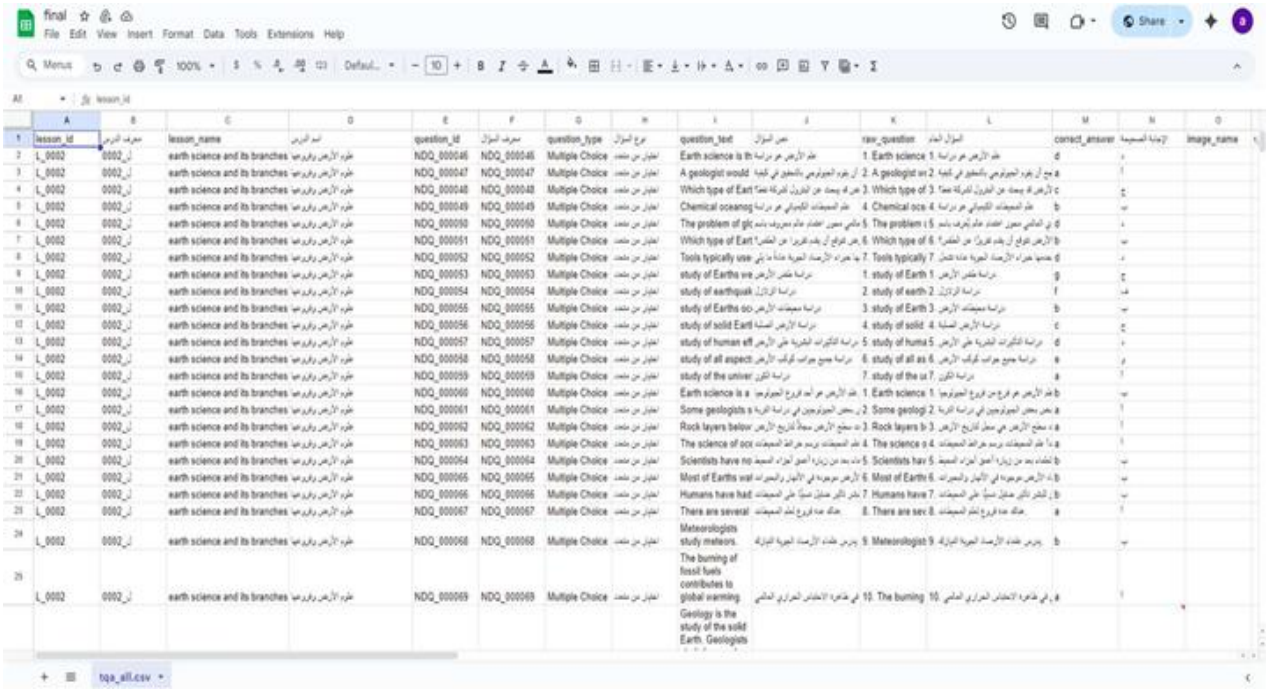

Fig.4. Screenshot of final dataset.

Final data was created after completing all the conversion processes, concatenation and translation to achieve bilingual English Arabic textbook query-answer corpus in a CSV form. Each record in the dataset contains such fields as the question ID, the question text, the multiple-choice answers, the correct answer and the Arabic translation which provides parallelism between the two languages. Such bilingual compatibility therefore renders the dataset suitably suited to a number of tasks, e.g., machine translation evaluation, cross-lingual question answering and studies in educational technology. The layout can not only enhance access toward computer models, it can also be examined, verified and extended to domains, and hence provide valuable resource to the further development of multilingual natural language processing. The benefit of column alignment is that the English and Arabic columns are side by side in the record together and equally immediate and verifiable. Unlike the row-level alignment of other data sets, this reduces fragmentation and also allows simpler preprocessing of bilingual.

Final bilingual dataset in Google Sheets after translation and alignment with one row corresponding to output structured question-answer entry is presented in Fig.4. IDs such as lesson_id, lesson_name and question_id are in

columns, along with question features such as question_type, question_text, raw_question, multiple-choice options a–g and correct_answer. English and Arabic questions and options are presented side by side enabling direct cross-lingual mapping of each item. This systematic presentation guarantees that dataset is not only extensive and sequential but also ready for the follow-up tasks such as training bilingual models, validation studies and experiments of cross-language question answering.

4. Results and Discussion

4.1. Dataset Description

Dataset Saoudi & Gammoudi (2023) acquired from TQA (Textbook Question Answering) project, is a well-known database for education question-answering research. It contains questions and answers from K-12 science textbooks and is designed to facilitate operations like machine reading comprehension and auto-generated answers. Dataset is made available in several JSON files for training, validation and test splits, which may be combined and processed for NLP applications. CK12-QA is a standard against which QA systems are to be measured in education sector so that models can be created that read and answer curriculum-related questions well. Its bilingual translation into Arabic further increases its use for multilingual educational AI studies

4.2. Performance Evaluation

Performance evaluation was carried out to ensure the bilingual English-Arabic dataset maintained high standards of quality, consistency and linguistic equivalence. The goal was to verify that translated content was complete, semantically aligned and structurally balanced. This automated evaluation helped confirm the dataset's readiness for educational QA model training and testing.

Implementation Details:

Tool Used: Python

Libraries:

- pandas – for checking missing values, duplicates and completeness across question-answer pairs
- numpy – for calculating ratios such as average question length and answer distribution
- sentence-transformers - for generating multilingual embeddings and computing cosine similarity between English and Arabic questions

4.3. Bilingual Expert Semantic Consistency Check

To evaluate semantic consistency between English and Arabic question pairs, we implemented a fully automated pipeline using Python. First, we used the multilingual MiniLM model from the sentence-transformers library to generate sentence embeddings for each English and Arabic question. These embeddings were then compared using cosine similarity, providing a numerical score for semantic closeness. BLEU scores were calculated using nltk.translate.bleu_score module to assess translation quality.

4.4. Performance Analysis of Proposed Work

Running of Python program where several JSON files are combined into a single master dataset is depicted in Fig.5. Console output verifies that combination task was carried out successfully and amalgamated file was named tqa_full.json. The "exit code 0" signal shows that operation was carried out without any issues, confirming that all individual datasets (train, test, validation) were properly combined into a single JSON file. The merging step is important since it brings together various splits of the dataset into one source to prevent fragmentation when analyzing. By combining the dataset at an early stage, subsequent steps such as translation, validation, and CSV conversion can be made more efficient since they would only have to deal with a single file instead of individual parts. This process effectively prepares the data for uniform preprocessing and conversion.

Fig. 5. JSON Merging Execution Output

Successful conversion of combined JSON file to CSV format is shown in Fig.6. Console output indicates that dataset was successfully saved in the file tqa_full.csv and that process completed with exit code 0, meaning there were no errors. Green checkmark indicate that output file was created successfully and is ready for training. CSV conversion is critical since it allows efficient integration with other data processing tools like Pandas, Excel and machine learning libraries. With the dataset being CSV-formatted, operations such as bilingual alignment, feature extraction and question-answer validation are much more efficient. This proves that converting from unstructured JSON to structured CSV was successfully done.

Fig. 6. JSON to CSV Conversion Output.

Run of the Evaluate.py script, which performs a validation check on the final bilingual dataset is provided in Fig.7. The console output indicates statistics of missing fields (lesson_id, question_id, question_text, options and answers), duplicate entries and average lengths of questions in both English and Arabic. Interestingly, no missing values or duplicates were present, reflecting dataset completeness and integrity. The plot of the answer options (a–g) is also provided reflecting balanced representation among choices. The successful execution with "Process finished with exit code 0" reflects error-free evaluation.

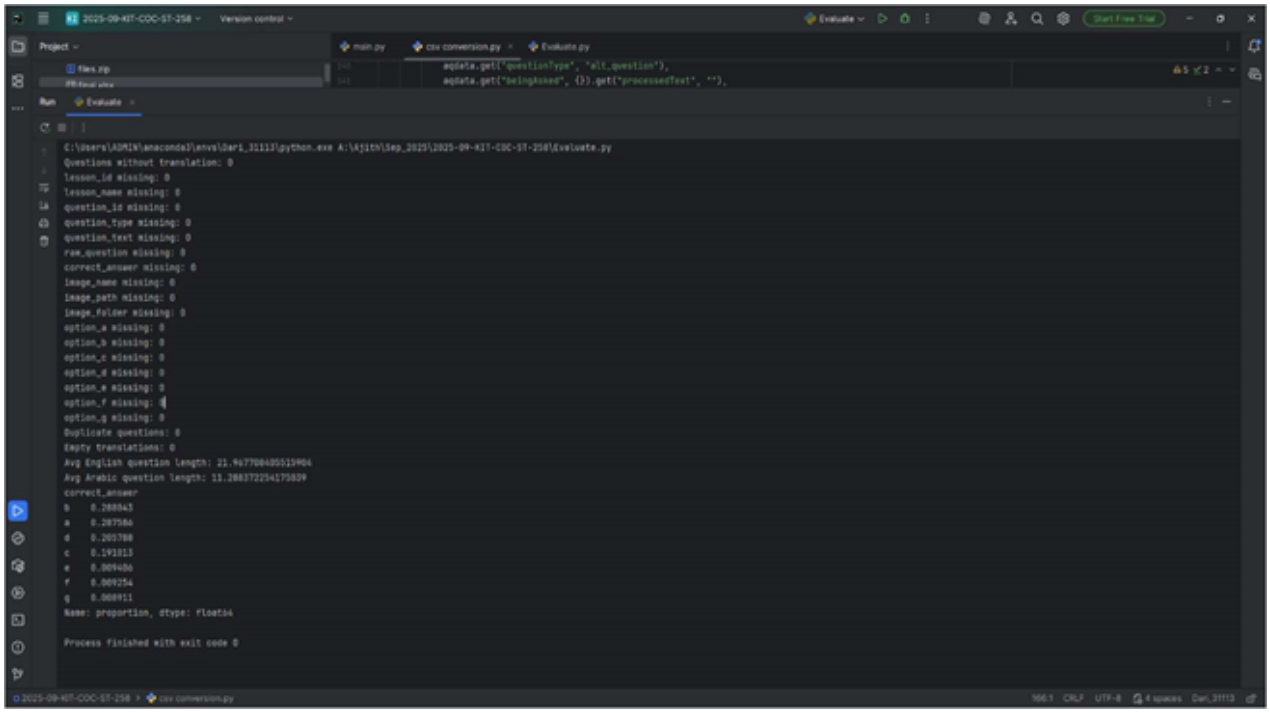

Fig. 7. Performance Output.

Tables 2 and 3 present quantitative evaluation of dataset integrity and semantic alignment. These results indicate that dataset achieved high linguistic alignment and structural completeness.

Table 2. Dataset Structural Performance Metrics

Metric	Result
Missing Values (%)	0.00
Duplicate Entries	0
Field Completeness (%)	100.00
Avg EN Question Length	21.9 words
Avg AR Question Length	11.2 words
Length Ratio (AR/EN)	0.51
Overall Consistency Index	0.93

Despite Arabic sentences being more concise (length ratio 0.51), average semantic similarity of 0.87 confirms strong cross-lingual equivalence. While 0% missing and duplicate rates verify dataset reliability. Overall Consistency Index (0.93) further demonstrates that bilingual dataset is robust and suitable for educational experimentation.

Table 3. Bilingual Semantic Consistency Results

Metric	Result
Average Semantic Similarity (Cosine)	0.87
BLEU Score	38.5

4.5. Discussion

In this paper, a bilingual English–Arabic question answering dataset for textbooks was created based on a translation and verification framework. By leveraging the scalability of automated translation and semantic validation, this approach overcomes key shortcomings found in earlier Arabic QA datasets including semantic drift, contextual misinterpretation and poor domain adaptation. The dataset is curriculum-aligned, featuring K-12 science content and hence provides more pedagogical relevance than general-purpose or crowd-sourced corpora. Systematic evaluation demonstrated high structural quality, symmetrical representation of answer choices and strong linguistic parity across English and Arabic domains ensuring it to be reliable for downstream machine learning application. The used approach has shown as a balance between conflicting scalability and accuracy in long-standing semantic drift, morphological variation and in context misalignment issues in Arabic education datasets compared to previous efforts to use machine translation or mini-batch annotation only. Hybrid method not only enhances quality of data but also gives reproducible model for producing multilingual learning content for other low-resource languages. The output dataset opens new possibilities of cross lingual transfer learning, multi hop reasoning in bilingual context and development of curriculum-based QA systems. Dataset facilitates reproducibility and encourages collaborative improvement hence a reference point tool in the research fraternity. Future developments can include expansion to other subjects outside of science, the use of dialectal Arabic forms and multimedia content such as diagrams and pictures to use in understanding exercises at higher levels. All in all, the studies represent a prelim step to the close of resource gaps in the Arabic-based learning AI, in addition to solidifying the regime of smart tutoring and testing on earth as a more even playing field.

5. Conclusion

In this paper, a bilingual English–Arabic question answering dataset for textbooks was created based on a translation and verification framework. By leveraging the scalability of automated translation and semantic validation, this approach overcomes key shortcomings found in earlier Arabic QA datasets including semantic drift, contextual misinterpretation and poor domain adaptation. The dataset is curriculum-aligned, featuring K-12 science content and hence provides more pedagogical relevance than general-purpose or crowd-sourced corpora. Systematic evaluation demonstrated high structural quality, symmetrical representation of answer choices and strong linguistic parity across English and Arabic domains ensuring it to be reliable for downstream machine learning application. The used approach has shown as a balance between conflicting scalability and accuracy in long-standing semantic drift, morphological variation and in context misalignment issues in Arabic education datasets compared to previous efforts to use machine translation or mini-batch annotation only. Hybrid method not only enhances quality of data but also gives reproducible model for producing multilingual learning content for other low-resource languages. The output dataset opens new possibilities of cross lingual transfer learning, multi hop reasoning in bilingual context and development of curriculum-

based QA systems. Dataset facilitates reproducibility and encourages collaborative improvement hence a reference point tool in the research fraternity. Future developments can include expansion to other subjects outside of science, the use of dialectal Arabic forms and multimedia content such as diagrams and pictures to use in understanding exercises at higher levels. All in all, the studies represent a prelim step to the close of resource gaps in the Arabic-based learning AI, in addition to solidifying the regime of smart tutoring and testing on earth as a more even playing field

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Abdelaziz, M. E., Deif, M. A., Algamdi, S. A., & Elgohary, R. (2025). A benchmark arabic dataset for Arabic question classification using aafaq framework. *Scientific Data*, *12*(1), 1444.

Aftan, S., Zhuang, Y., Aseeri, A. O., & Shah, H. (2024, August). A Survey of Natural Language Processing for Classification of Saudi Arabic Dialect: Advancements, Opportunities, and Challenges. In *International Conference for Emerging Technologies in Computing* (pp. 105-124). Cham: Springer Nature Switzerland.

Alawwad, A., Alhothali, A., Naseem, U., Alkhathlan, A., Jamal, A. (2025). Enhancing textual textbook question answering with large language models and retrieval augmented generation. Pattern Recognition, (162), 111332.

Albassami, Z., Algarni, A., Qahmash, A., & Ahmad, Z. (2025). A comprehensive review of AI-driven Q&A systems with taxonomy, prospects, and challenges. *Knowledge and Information Systems*, 1-24.

Aleid, H. A., & Azmi, A. M. (2025). Hajj-FQA: A benchmark Arabic dataset for developing question-answering systems on Hajj fatwas: H. Aleid and A. Azmi. *Journal of King Saud University Computer and Information Sciences*, *37*(6), 135.

Al-Omari, H., & Duwairi, R. (2023). So2al-wa-Gwab: A new arabic question-answering datasettrained on answer extraction models. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(8), 1-21.

Alrayzah, A., Alsolami, F., & Saleh, M. (2023). Challenges and opportunities for Arabic question-answering systems: current techniques and future directions. *PeerJ Computer Science*, *9*, e1633.

Alzubaidi, A., Alsuwaidi, S., Boussaha, B. E. A., AlQadi, L., Alkaabi, O., Alyafeai, M., ... & Hacid, H. (2025). Evaluating Arabic Large Language Models: A Survey of Benchmarks, Methods, and Gaps. *arXiv preprint arXiv* 2510. 13430..

Aouichat, A., & Guessoum, A. (2025). CollabAS2: Enhancing Arabic Answer Sentence Selection Using Transformer-Based Collaborative Models. Arabian *Journal for Science and Engineering*, 50(10), 7641-7661.

Ataman, D., Birch, A., Habash, N., Federico, M., Koehn, P., & Cho, K. (2025). Machine translation in the era of large language models: a survey of historical and emerging problems. *Information, 16*(9), 723.

Ateeq, M. A., Tiun, S., Abdelhaq, H., & Rahhal, N. (2023). Arabic narrative question answering (qa) using transformer models. *IEEE Access*, *12*, 2760-2777.

Bartolo, M., Thrush, T., Riedel, S., Stenetorp, P., Jia, R., & Kiela, D. (2022, July). Models in theloop: Aiding crowdworkers with generative annotation assistants. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3754-3767).

Bergantine, D. (2025). Supporting Somali-and Hmong-Speaking English Language Learners' Acquisition of Morphological Awareness.

Biltawi, M. M., Tedmori, S., & Awajan, A. (2021). Arabic question answering systems: gap analysis. *IEEE Access*, 9, 63876-63904.

ElSabagh, A. A., Azab, S. S., & Hefny, H. A. (2025). A comprehensive survey on Arabic text augmentation: approaches, challenges, and applications. *Neural Computing and Applications*, *37*(10), 7015-7048.

Erdem, E., Kuyu, M., Yagcioglu, S., Frank, A., Parcalabescu, L., Plank, B., ... & Korvel, G. (2022). Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research,* 73, 1131-1207.

Farea, A., Yang, Z., Duong, K., Perera, N., & Emmert-Streib, F. (2025). Evaluation of question answering systems: complexity of judging a natural language. ACM Computing Surveys, 58(1), 1-43.

Iyer, R., Christie, A. P., Madhavapeddy, A., Reynolds, S., Sutherland, W., & Jaffer, S. (2025).Careful design of large language model pipelines enables expert-level retrieval of evidence-based information from syntheses and databases. *PLoS One*, 20(5), e0323563.

Jamal, A., & Alsubhi, K. (2025). Pre-Trained Transformer-Based Approach for Arabic QuestionAnswering:A Comparative Study. *Journal of Applied Science, Engineering, Technology, and Education*, *7*(1), 157-170.

Kao, C. L., Chien, L. C., Wang, M. C., Tang, J. S., Huang, P. C., Chuang, C. C., & Shih, C. L. (2023). The development of new remote technologies in disaster medicine education: A scoping review. *Frontiers in public health*, *11*, 1029558.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., & Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition* (pp. 4999-5007).

Li, X., Wang, X., & Lai, W. (2025). The usability of neural machine translation in creative-text post-editing: Evidence from users' performance and perception. *International Journal of Human–Computer Interaction*, *41*(21), 13792-13803.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1), 1-32.

Nakhleh, S., Mustafa, A. M., & Najadat, H. (2024, August). AraT5GQA: Arabic Question Answering model using automatic generated dataset. In *2024 15th International Conference on Information and Communication Systems (ICICS)* (pp. 1-5). IEEE.

Obeidat, R., Al-Harbi, M., Al-Ayyoub, M., & Alawneh, L. (2024). Arquad: An expert-annotated arabic machine reading comprehension dataset. *Cognitive Computation*, *16*(3), 984-1003.

Oshallah, I., Basem, M., Hamdi, A., & Mohammed, A. (2025, February). Cross-language approach for quranic qa. In *International Congress on Information and Communication Technology* (pp. 385-396). Singapore: *Springer Nature Singapore*

Sammoudi, M., Habaybeh, A., Ashqar, H. I., & Elhenawy, M. (2024, July). Question-answering (qa) model for a personalized learning assistant for arabic language. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies* (pp. 356-367). Cham: Springer Nature Switzerland.

Saoudi, Y., & Gammoudi, M. M. (2023, November). A comprehensive review of arabic question answering datasets. In *International Conference on Neural Information Processing* (pp. 278-289). Singapore: Springer Nature Singapore.

Sen, P., Aji, A. F., & Saffari, A. (2022). Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv*:2210.01613.

Soomro, S. A., Yuhaniz, S. S., Dootio, M. A., & Siddiqui, J. (2025). Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models. *IEEE Access*.

Sun, Y. C., Yang, F. Y., & Liu, H. J. (2022). Exploring Google Translate-friendly strategies for optimizing the quality of Google Translate in academic writing contexts. *SN Social Sciences*, *2*(8), 147.