PersianMCQ-Instruct: A Comprehensive Resource for Generating Multiple-Choice Questions in Persian

Kamyar Zeinalipour[1], Neda Jamshidi[1], Fahimeh Akbari[1], Marco Maggini[1], Monica Bianchini[1], Marco Gori[1],

[1]University of Siena, DIISM, Via Roma 56, 53100 Siena, Italy

Correspondence: kamyar.zeinalipour2@unisi.it

Abstract

We present *PersianMCQ-Instruct*[1 2], a comprehensive resource comprising a dataset and advanced models for generating multiple-choice questions (MCQs) in standard Iranian Persian, a low-resource language spoken by over 80 million people. This resource includes three state-of-the-art models for Persian MCQ generation: *PMCQ-Gemma2-9b*[3], *PMCQ-Llama3.1-8b*[4], and *PMCQ-Mistral-7B*[5]. Inspired by the Agent Instruct framework and *GPT-4o*, we created the dataset by curating over 4,000 unique Persian Wikipedia pages, generating three MCQs per page for a total of over 12,000 questions. To ensure the quality of the dataset, we conducted both human evaluations and model fine-tuning, which showed substantial performance improvements in the Persian MCQ generation. The dataset and models are publicly available, providing valuable tools for researchers and educators, with a particular impact on enhancing Persian-language educational technology.

1 Introduction

Generating high-quality multiple-choice questions is essential for various applications, including educational assessments, language learning tools, and automated tutoring systems. These questions efficiently evaluate comprehension and knowledge retention.

In natural language processing (NLP), large language models (LLMs) have significantly enhanced automated text generation and comprehension. This paper introduces a novel method for generating Persian multiple-choice questions (MCQs) by fine-tuning LLMs. Inspired by the Agent Instruct framework (Mitra et al., 2024), We developed a high-quality Persian Multiple Choice Question (MCQ) dataset named *PersianMCQ-Instruct* to address the shortage of educational resources for the Persian language, a low-resource language in the field of NLP. This dataset is crafted to meet rigorous educational standards and supports the growing need for Persian language resources in educational technology.

PersianMCQ-Instruct, which includes corresponding texts sourced from prominent Persian-language Wikipedia pages. We then fine-tuned the LLM using this dataset.

In this study, we significantly contribute to Persian MCQ generation by introducing the *PersianMCQ-Instruct* dataset, a comprehensive collection of Farsi articles from WikiFarsi paired with well-designed MCQs. This pioneering resource enables the development and evaluation of MCQ generation models tailored specifically to the Persian language.

Drawing from the aforementioned framework, we employ a three-step approach for MCQ generation from text: content transformation, MCQ generation, and MCQ refinement. Each phase ensures high-quality educational questions. We used the *GPT-4o* model, resulting in the high-quality *PersianMCQ-Instruct* dataset.

To demonstrate *PersianMCQ-Instruct* quality and enhance Persian MCQ generation, we fine-tuned several LLMs*gemma2-9b-it*, *Llama3.1-8b-Instruct*, and *Mistral-7b-Instruct-v0.3*using *PersianMCQ-Instruct* dataset. This resulted in improved models: *PMCQ-Gemma2-9b*, *PMCQ-Llama3.1-8b*, and *PMCQ-Mistral-7b*, designed to automate Persian MCQ and answer generation from educational content.

[1]https://huggingface.co/datasets/Kamyar-zeinalipour/PersianMCQ-instruct

[2]https://github.com/KamyarZeinalipour/Persian_MCQ

[3]https://huggingface.co/Kamyar-zeinalipour/PMCQ-Gemma2-9b

[4]https://huggingface.co/Kamyar-zeinalipour/PMCQ-Llama3.1-8b

[5]https://huggingface.co/Kamyar-zeinalipour/PMCQ-Mistral-7B

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 344–372
January 20, 2025.

We evaluated the models' performance before and after fine-tuning using the *PersianMCQ-Instruct* dataset, demonstrating marked improvements and underscoring the effectiveness of the fine-tuning process.
Finally, we made the *PersianMCQ-Instruct* dataset and all fine-tuned models publicly available to foster further research and practical application, advancing Persian MCQ generation and enhancing educational resources.
The paper is structured as follows: Section 2 reviews relevant literature. Section 3 details our methodology. Section 4 analyzes experimental results. Section 5 concludes and Section 6 discusses the limitations of this study.

2 Related Work

Question Generation (QG) is a crucial task in natural language processing that involves automatically creating questions from a given sentence or paragraph. This task is challenging as it requires identifying key statements within the context and generating questions based on them. QG can be categorized into answer-aware and answer-agnostic types (Dugan et al., 2022). Multiple-choice questions (MCQs) are particularly important in educational settings, where they are widely used in assessments and exams. For languages like Persian, where resources are limited, developing MCQs is especially crucial to enhance educational tools and materials.

QG and Question Answering (QA) are interconnected tasks that require reasoning between questions and answers. Datasets originally created for QA tasks, such as SciQ, RACE, and FairytaleQA (Welbl et al., 2017; Lai et al., 2017; Xu et al., 2022), are also used in QG research (Tang et al., 2017; Jia et al., 2021; Steuer et al., 2022; Zhao et al., 2022). Specialized datasets for QG include LearningQ, KHANQ, and EduQG, which cover various subjects and educational levels (Chen et al., 2018; Gong et al., 2022; Hadifar et al., 2023a).

Early QG methods relied on rule matching, but advancements have led to the use of Seq2Seq models with attention, linguistic feature integration, multi-modal models, multi-task learning, reinforcement learning, and language models like BERT and GPT-3 (Du and Cardie, 2017; Harrison and Walker, 2018; Zhou et al., 2018; Naeiji, 2022; Wang and Baraniuk, 2023; Zhou et al., 2019; Chen et al., 2019; Chan and Fan, 2019; Wang et al., 2022b; Sun et al., 2018; Yuan et al., 2017; Ma et al., 2020). Researchers have proposed encoding answers with context, utilizing answer positions, or using text summaries to incorporate answer information in both answer-aware and answer-agnostic QG (Sun et al., 2018; Yuan et al., 2017; Ma et al., 2020).
AgentInstruct is an extensible agentic framework designed to automatically generate large volumes of diverse and high-quality synthetic data. This framework leverages raw data sources such as text documents and code files as seeds to create both prompts and responses. The process involves three main stages: Content Transformation Flow, where raw text is transformed into structured content like argument passages or API lists; Seed Instruction Generation Flow, where this transformed content is used to generate a comprehensive set of seed instructions; and Instruction Refinement Flow, where these instructions are iteratively refined to enhance quality, diversity, and complexity. By automating these steps, AgentInstruct aims to facilitate Generative Teaching, enabling powerful models to teach new skills or behaviors to other models efficiently (Mitra et al., 2024). This framework can significantly aid in generating multiple-choice questions by creating a wide range of high-quality questions and options, thereby enhancing the training data for MCQ generation models.
In educational contexts, controlling question difficulty is crucial for effective education, with methods assessing difficulty based on answerability, inference steps needed, or learners' abilities (Lord, 2012; Qiu et al., 2020; Uto et al., 2023). Aligning questions with the syllabus is important for test focus, leading to studies training classifiers or ranking models to determine question relevance (Hadifar et al., 2023b). Personalized education requires generating customized questions for students, prompting the development of knowledge-tracking models based on student answer histories or few-shot knowledge-tracking models incorporating sequences of student states and questions (Wang et al., 2023; Srivastava and Goodman, 2021). Previous works in developing educational tools with LLMs have also focused on this aspect (Zeinalipour et al., 2023a),(Zeinalipour et al., 2023b) and (Zeinalipour et al., 2023c). Significant advancements have been achieved in educational technology for specific languages. Notably, a Turkish MCQs generator has been successfully developed (Zeinalipour et al., 2024b). Additionally, Inspired by self-instruct methods, several works

have explored various languages, including Turkish, Arabic, English, and Italian. (Zeinalipour et al., 2024c) ,(Zugarini et al., 2024) and (Zeinalipour et al., 2024a).

In the realm of Persian natural language processing, notable works include ParsiNLU, which covers challenges in reading comprehension, multiple-choice question-answering, textual entailment, sentiment analysis, question paraphrasing, and machine translation (Khashabi et al., 2021). PersianMind achieved state-of-the-art results on the Persian subset of the Belebele benchmark and the ParsiNLU multiple-choice QA task (Rostami et al., 2024). Efforts to develop QA systems for Persian have involved translating English QA datasets, but this approach often fails to produce high-quality annotated data due to translation imperfections. There are few open-domain QA datasets for Persian. For instance, Abadani et al. (Abadani et al., 2021) translated SQuAD into Persian, creating ParSQuAD, and Kazemi et al. (Kazemi et al., 2022) developed PersianQuAD, a native dataset where annotators pose questions and specify answers within paragraphs. Despite these advancements, there remains a critical gap in the generation of text augmentation and multiple-choice questions with answers in Persian. Existing works have not adequately addressed the need for comprehensive and high-quality resources in this area. To fill this void, our study introduces the *PersianMCQ-Instruct* dataset, specifically designed for generating multiple-choice questions in Persian. Furthermore, we present several fine-tuned models, including *PMCQ-Gemma2-9b*, *PMCQ-Llama3.1-8b*, and *PMCQ-Mistral-7b*, tailored for generating MCQs from text in Persian. This work is essential for advancing natural language processing in the Persian language, offering valuable resources for educational applications and language assessment tools. By addressing the current limitations, our contributions aim to significantly enhance the quality and availability of Persian MCQ datasets, thereby fostering further research and development in this field.

3 Methodology

In this study, we introduce the development of an advanced Persian educational multiple-choice question (MCQ) generator, leveraging state-of-the-art large language models (LLMs). We have curated an extensive dataset, named *PersianMCQ-Instruct*, which includes multiple-choice questions derived from Persian texts. To generate and evaluate Persian MCQs utilizing the *PersianMCQ-Instruct* dataset, we fine-tuned a variety of LLMs across multiple scenarios, focusing on the multiple-choice format. The models optimized in this process included *Llama3.1-8b-Instruct*, *gemma2-9b-it*, and *Mistral-7b-Instruct-v0.3*. This section outlines the methodologies employed for dataset generation and model fine-tuning, providing a detailed account of the procedures followed to establish an effective Persian MCQ generator. Figure 1 presents the comprehensive methodology applied in this study.

3.1 PersianMCQ-Instruct

In the preceding sections, we presented a comprehensive dataset focused on Persian educational materials, encompassing texts from various academic disciplines. This dataset includes multiple-choice questions. The creation process, illustrated in Figure 1, involved scraping content from various online sources, including Wikipedia, followed by data cleaning, filtering, and the design of prompts inspired by Agent Instruct for generative processes. The questions and answers were generated using *GPT-4o*, renowned for its superior natural language understanding. This step was pivotal in producing realistic and challenging questions. We conducted an exhaustive evaluation to ensure the accuracy, relevance, and educational utility of the questions. This analysis seeks to demonstrate how this dataset enhances educational resources in Persian and can serve as a model for analogous initiatives in other languages and disciplines.

3.1.1 Data Scraping

The process of information extraction starts with a focused filtering phase aimed at specific Persian Wikipedia pages known for presenting widely consumed content across diverse academic fields. This comprehensive repository includes materials spanning a multitude of subjects such as mathematics, history, biology, and literature. For this research, we have assembled a dataset derived from a variety of primary Wikipedia online resources.[6] [7] [8] [9] [10] [11]

[6] List of most viewed articles by topic.
[7] Offline version project.
[8] Featured articles.
[9] Good articles.
[10] 100 essential articles.
[11] Vital articles level 2.

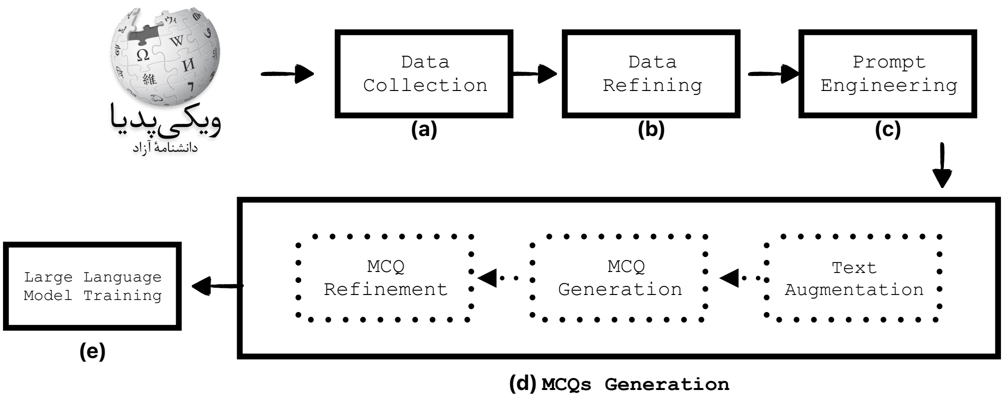

Figure 1: The figure illustrates the methodology employed in this study, which includes the following steps: (a) Data collection by scraping content from popular Persian Wikipedia pages. (b) Data refinement and filtering to enhance quality by eliminating overly short or excessively detailed content. (c) Creation of prompts for generating Persian multiple-choice questions (MCQs) based on the refined text. (d) Utilization of *GPT-4o* to produce quizzes from the gathered data and configured prompts, involving three sub-steps: (i) text augmentation, (ii) MCQ generation, and (iii) MCQ refinement. (e) Fine-tuning of large language models (LLMs) with the generated dataset to generate Persian MCQs from the provided context.utilizing models *gemma2-9b-it*,*Llama3.1-8b-Instruct*, and *Mistral-7b-Instruct-v0.3*.

[12] [13] [14] [15] These sources offer in-depth summaries of essential concepts and topics that are aligned with educational content.

3.1.2 Data Cleaning

After scraping Wikipedia, we initially gathered 8,894 pages. However, to ensure the quality and coherence of the content, we applied several filtering criteria. First, we removed pages with less than 100 words, as these often lacked sufficient information for generating meaningful MCQs. The number of examples exceeding 500 words was quite small. Due to our resource limitations during the training phase, we decided to cap the examples to 500 words. We aimed to maintain a consistent word count range for our data set.
Additionally, we conducted a review to identify and remove pages containing sensitive content to align with our ethical considerations. Following this data filtering process, we were left with 4,159 pages, which then served as the basis for generating MCQs in Persian.

[12] Vital articles.
[13] Essential articles every Wikipedia should have.
[14] List of lists of lists.
[15] Specialized articles needed.

3.1.3 Craft the prompts

Crafting targeted prompts was a crucial aspect of our methodology. As previously mentioned, we drew inspiration from *Agent Instruct* to generate our MCQ dataset in Persian. The generation process involved the following steps: 1. **Text Augmentation**: We first augmented the provided text from Wikipedia to enhance data quality and improve performance, which is particularly valuable for low-resource languages, to prepare it for the generation process. 2. **MCQ Generation**: Using the augmented text from the initial step, we proceeded to generate MCQs. 3. **MCQ Refinement**: We then refined the generated MCQs by incorporating both the augmented text from the first step and the initially generated MCQs. To accomplish these tasks, we created three distinct prompts and applied prompt engineering techniques, experimenting with various prompts to optimize each step. The specific prompts used at each stage are illustrated in Figure 2, and all prompts have also been included in the appendix A.

3.1.4 Generating Persian MCQs.

Our methodology utilizes Large Language Models (LLMs) to autonomously generate three Persian multiple-choice questions (MCQs) for each text, inspired by the Agent Instruct framework principles.

```
As an Educational Assistant your role is to transform
the given Persian text into an argumentative passage,
supporting a specific viewpoint and main topic with
relevant evidence. Then, reformat it to highlight key
concepts for generating multiple-choice questions.
1. Transform the original text into an argumentative
passage.
2. Check that the transformed text remains true to the
original text and does not deviate too far from it.
3. Ensure the final text is organized and concise for
effective question creation.
4. Proofread for grammar and spelling.
5. Ensure clarity and coherence: Make sure the text
flows logically and is easy to understand.
6. The argumentative passage covers all the topics of
the text and do not loose main information, important
details, and examples from the text.
The output should be a concise well-structured text in
Persian.I just need the argumentative passage as an
output, nothing else.
\n\n TEXT: {original_text}
```

(a) Text augmentation Prompt

```
As an Educational Assistant, your role is to generate multiple-choice
questions in Persian from the given text. Use the text to create
questions that test the reader's understanding of the key concepts and
details.
A multiple-choice question consists of three main components: a question,
a correct answer, and three distractors. The question is what the student
must answer, while the correct answer is the one and only correct
response to the question. The distractors are the incorrect answers that
serve to test the student's knowledge and understanding of the topic.
1. Identify key concepts and details in the text.
2. Create clear, concise, and relevant multiple-choice questions and the
question exists directly in the text.
3. Ensure each question has four answer choices, with one correct answer
and three distractors.
4. Make sure the correct answer exists directly in the text and is true.
5. Make sure the question is grammatically correct in Persian and is easy
for students to understand and should not be overly complex or ambiguous.
6. The distractors should be plausible but incorrect answers that are
related to the topic and the question. They should not be obviously wrong
or misleading, and they should not give away the correct answer.
7. Avoid using overly long sentences in the question, correct answer, or
distractors.
8. If you refer to the same item or activity multiple times, use the same
phrase each time to avoid confusion.
9. Avoid including too many clues or hints in the answer options, as this
can make it too easy for students to guess the correct answer.
10. Ensure that the questions are identical and not repetitive, even in
content, to maintain variety and challenge.
11. Generate three questions from the text that maintain their quality,
cover all the main topics, and are accurate.

Each string in the list should follow this format:
1. The first line contains the main question text.
2. The subsequent lines contain the answer options, each starting with
one of the letters الف, ب, ج, or د, followed by a period, the text of the
option, and then an indicator in parentheses specifying whether it is the
correct answer (پاسخ صحیح) or an incorrect answer (پاسخ غلط).
3. Correct questions should always be the first choice in الف.

Here is an example with placeholders:

[
    "Question text 1?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect
answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد. Incorrect answer
(پاسخ غلط)",
    "Question text 2?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect
answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد. Incorrect answer
(پاسخ غلط)",
    "Question text 3?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect
answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد. Incorrect answer
(پاسخ غلط)"
]

Do not include 'in' or 'sider' when generating questions; use them only
for determining list members.

\n\nTEXT: {first_step_text}
```

(b) MCQ generation Prompt

```
As an Educational Assistant, your role is to increase the
complexity of the three multiple-choice questions that you
receive as input, using the text for this purpose.
1. For each question, make the structure of question more
complex, requiring careful analysis and deeper thought to
understand it and it exists directly in  the text.
 2. Ensure that the correct answer and the three distractors
are more nuanced and closely related, requiring careful
analysis to discern the correct option.
3. The correct answer and three distractors should be more
complex in content and structure, and the correct answer
should be directly in the text and be true.
4. Incorporate common misconceptions or similar terminology to
test deeper understanding.
5. Ensure that the distractors are plausible but falls  and
subtle enough to challenge the reader's comprehension.
6. Make sure the questions and answers are  grammatically
correct in Persian.
7. The questions and answers should not be  long when compared
with the original questions.

Each string in the list should follow this format:
1. The first line contains the main question text.
2. The subsequent lines contain the answer options, each
starting with one of the letters الف, ب, ج, or د, followed by
a period, the text of the option, and then an indicator in
parentheses specifying whether it is the correct answer (پاسخ
صحیح) or an incorrect answer (پاسخ غلط).
3. Correct questions should always be the first choice in الف.
Here is an example with placeholders:

[
    "Question text 1?\nالف. Correct answer (پاسخ صحیح)\nب.
Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)
\nد. Incorrect answer (پاسخ غلط)",
    "Question text 2?\nالف. Correct answer (پاسخ صحیح)\nب.
Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)
\nد. Incorrect answer (پاسخ غلط)",
    "Question text 3?\nالف. Correct answer (پاسخ صحیح)\nب.
Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)
\nد. Incorrect answer (پاسخ غلط)"
]

Do not include 'in' or 'sider' when generating questions; use
them only for determining list members.

\n\n TEXT: {first_step_text} \n Multiple-Choice Questions:
{second_step_text}
```

(c) MCQ refinement Prompt

Figure 2: Three prompts which we used for Persian MCQ generation

Our approach integrates these generated questions with contextual inputs, ensuring relevance and coherence. Here's how we achieve high-quality Persian MCQs from a given text: We begin by enhancing the provided Wikipedia text for the generation process. Next, we use the augmented text to create the MCQs. Finally, we refine the MCQs by incorporating elements from both the augmented text and the initial MCQs. We employ *GPT-4o* for its efficiency and performance, leveraging meticulously curated content, topics, and prompts to produce tailored multiple-choice questions aligned with educational goals. Figure 3 presents the word distribution across the generated Persian multiple-choice questions (MCQs) and their corresponding contexts. In the upper plot, the X-axis denotes the word distribution of the used contexts, while the Y-axis displays the number of utilized Wikipedia pages. In the lower plot, the X-axis maintains the word distribution of generated Persian MCQs, with the Y-axis showing the number of generated Persian MCQs.

3.1.5 Evaluating *PersianMCQ-Instruct* Quality

Assessing the quality of generated Persian MCQs faces a significant challenge due to the absence of a reference corpus, which is essential for benchmarking these questions using metrics such as ROUGE scores (Lin, 2004). ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores measure the overlap between generated and reference summaries and are widely used for summarization quality evaluation. These scores include:

- **ROUGE-1**: Counts matching single words (unigrams), indicating core content similarity.
- **ROUGE-2**: Counts matching word pairs (bigrams), capturing content and flow.
- **ROUGE-L**: Finds the longest common subsequence, reflecting content and sentence structure without requiring consecutive matches.

This deficiency complicates the evaluation of educational MCQs, as creating effective questions requires subtle rewording of reference texts. An evaluation method focusing on high levels of textual extraction is necessary to address this unique challenge.
To overcome this obstacle, our methodology employs ROUGE-L scores to evaluate the degree to which the generated questions adhere to the original context. Encouragingly, the results from our approach have been promising. We achieved an average ROUGE-L F1 score of 0.022 and a BERTScore F1 of 0.89, demonstrating a strong correlation between the generated questions and the corresponding sentences in the source material. For detailed results on all ROUGE and BERT scores, please refer to Table 1.

Metric	Precision	Recall	F1
ROUGE-1	0.0167	0.0690	0.0238
ROUGE-2	0.0060	0.0246	0.0084
ROUGE-L	0.0158	0.0667	0.0226
BERTScore	0.8886	0.8976	0.8929

Table 1: Average ROUGE and BERTScore Results for Context and Generated MCQs

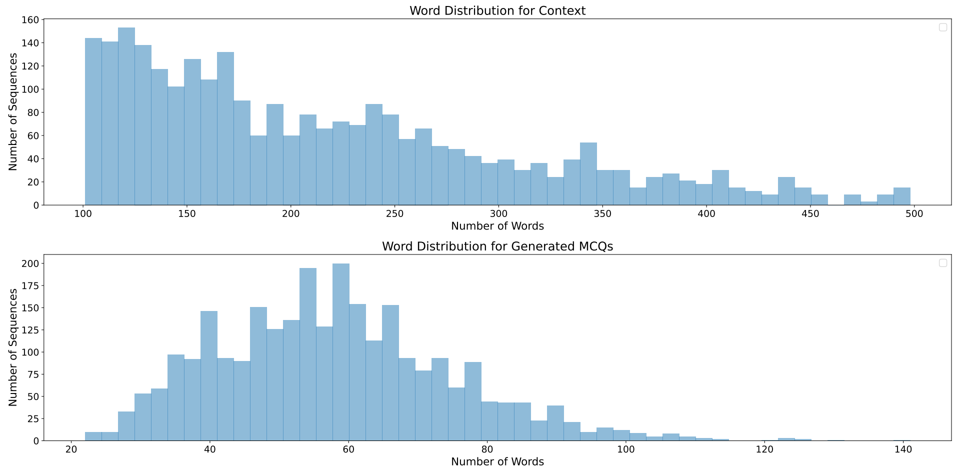

Figure 3: The word distribution in the generated Persian MCQs and their respective contexts

In addition to the quantitative metrics, a comprehensive qualitative evaluation was carried out using human assessors. We engaged native Persian speakers with deep linguistic expertise to evaluate the generated quizzes, ensuring a thorough and nuanced assessment process. This approach was designed to reflect the complexities involved in constructing and evaluating effective educational questions in Persian.

A selection of 600 questions was reviewed by a panel of Persian language experts. This panel consisted of two individuals whose native language is Persian and who are both postgraduate students at a university, possessing sufficient knowledge to evaluate the questions. Each expert evaluated 400 questions, with an overlap of 200 questions to assess annotator agreement. Inspired by the structured evaluation framework described in (Wang et al., 2022a), we implemented a rigorous five-point rating system for human assessment of the generated dataset. This rating scale facilitated a nuanced evaluation across five distinct levels, thereby ensuring that our models approximate human-like instruction-following behaviors without the necessity for extensive manual annotation. Our approach combined both quantitative and qualitative methods, enabling a comprehensive analysis of the generated multiple-choice questions (MCQs) in terms of their efficiency and effectiveness. The specifics of the five-point rating system adopted for our evaluation are outlined below:
1. `RATING-A`: Questions and answers are factually accurate and directly relate to significant concepts from the source text. They are meaningful and precise without any grammatical issues or missing words. 2. `RATING-B`: Questions and answers are mostly factual and related to the text, though they may have some minor grammatical issues or be incomplete. There may be some missing parts in the answers, but they are still meaningful, or the true/false answer is not explicitly pointed out or failure to identify the correct answers. 3. `RATING-C`: Questions or answers are loosely related to the text but may address topics tangentially or may have some serious grammatical issues or Answers are not correct. 4. `RATING-D`: Questions or answers contain factual inaccuracies or are minimally relevent to the text. 5. `RATING-E`: Questions or answers or both not generated, Questions are demonstrably wrong or misleading or have no clear connection to the educational text .
Figure 4 shows the results of this human evaluation. Impressively, 86.8% of the generated questions received a rating of `RATING-A`, indicating that the newly introduced dataset maintains a high standard of quality. The agreement rate between the annotators was 0.96 and the Cohen's kappa (Gerald Rau, 2021) was 0.86. In Appendix D, we discuss the Human Evaluation on *PersianMCQ-Instruct*, assessing the quality of the questions we generated and providing a detailed analysis. In Appendix C, you can find various examples labeled with explanations detailing the rationale behind the specific labeling provided for each example.The evaluated data and code for the human annotation user interface (UI) are available on GitHub[16].

[16] `https://github.com/KamyarZeinalipour/HumanAnnotation-UI-PMCQ`

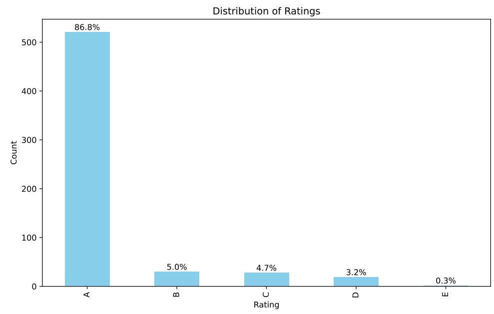

Figure 4: Distribution of human ratings for the MCQs generated by *PersianMCQ-Instruct*

3.2 From LLMs to Persian MCQs

To produce Persian multiple-choice questions from Persian textual resources and evaluate our innovative dataset, *PersianMCQ-Instruct*, we undertook a fine-tuning process involving various large language models, including *Llama3.1-8b-Instruct*[17], *Mistral-7b-Instruct-v0.3*[18], and *gemma2-9b-it*[19]. These models were chosen due to their comprehensive support for the Persian language. The fine-tuning procedure was extensive and meticulous, incorporating Parameter Efficient Fine-Tuning techniques to reduce task-specific loss. This rigorous approach aimed to not only deepen the models grasp of educational content but also ensure the accurate and nuanced generation of questions in Persian. Given the diverse nature of the content and the complexity of the language, achieving high fidelity in language generation was particularly challenging.

Before deploying these models for the task of question generation, they underwent significant customization through training specifically tailored to the under-investigated task. This customization phase was heavily supported by *PersianMCQ-Instruct*, a meticulously curated dataset introduced in a previous section. The dataset formed a critical basis for adapting the models, boosting their capability to formulate questions that are both contextually appropriate and linguistically precise in Persian.

4 Experiments

This section outlines the experimental procedures used to fine-tune Large Language Models (LLMs) to improve their performance in generating multiple-choice questions (MCQs) in Persian. The *PersianMCQ-Instruct* dataset, created as detailed in Section 3, served as the foundation for this process.

We employed this dataset to fine-tune three distinct LLMs: *Llama-3.1-8b-Instruct*, *Mistral-7b-Instruct-v0.3*, and *gemma2-9b-it*. The dataset was divided into two subsets for the fine-tuning phase. The first subset was allocated for training, consisting of 12,000 MCQs and 4,000 unique texts (as previously noted, we generated three different questions from each text).

The second subset, used for evaluation, consisted of 476 MCQs and 159 texts. This set was assessed using both automated metrics and human evaluation methods. The specific objective was to evaluate each model's capability to generate MCQs from Persian texts.

4.1 Training Setup

We leveraged two NVIDIA A6000 GPUs, each equipped with 48 GB of GPU RAM, for the training process. The training was conducted over 3 epochs with a maximum sequence length of 3500 tokens. A learning rate of 1e-4 was utilized, regulated by a cosine scheduler, along with a weight decay of 1e-4.

The batch size was maintained at 4 for both training and evaluation, and gradient accumulation was performed over 4 steps. Gradient checkpointing and flash attention (Dao, 2023) were enabled to optimize memory usage. Additionally, we employed LoRA (Hu et al., 2021) with a rank of 16 and an alpha of 32 to enhance the model performance. DeepSpeed (Rajbhandari et al., 2020) was used to improve computational efficiency and scalability. We used a dataset of 12,000 samples for training and 476 samples for evaluation.

4.2 Persian MCQs generation

To extract insights from specific multiple-choice questions in Persian text, we utilized the *PersianMCQ-Instruct* dataset, as detailed in Section 3. We fine-tuned several small-sized Large Language Models (LLMs) ranging from 7b to 9b parameters. Initially, these models had low performance in generating Persian questions. However, after fine-tuning using *PersianMCQ-Instruct* dataset, we observed significant improvements in output quality. These results confirm the high qual-

[17]Llama3 GitHub Repository
[18]Mistral GitHub Repository
[19]gemma2 GitHub Repository

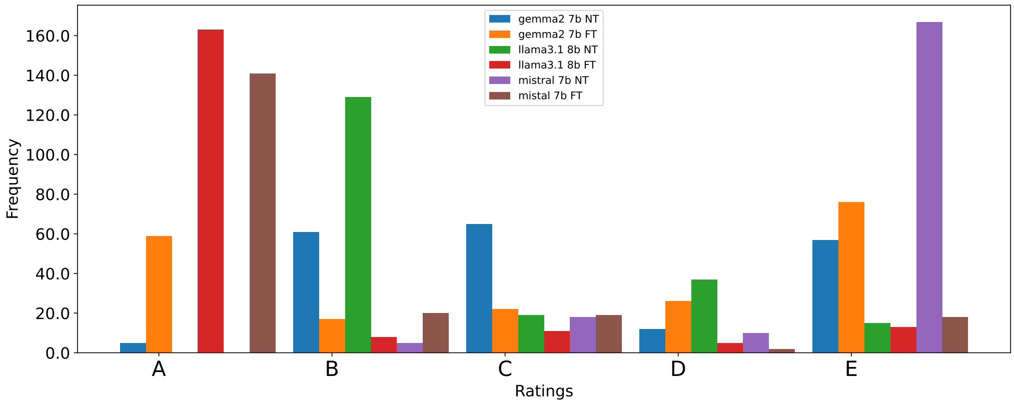

Figure 5: Human evaluation of the performance of LLMs on Persian MCQs with fine-tuning (FT) and without fine-tuning (NT).

	Model Name	**#param**	**Average Precision**	**Average Recall**	**Average F1**
Base	*gemma2-9b-it*	9	0.9004	0.8754	0.8877
	Llama3.1-8b-Instruct	8	0.8992	0.8630	0.8807
	Mistral-7b-Instruct-v0.3	7	0.8852	0.8631	0.8740
Fine-tuned	*PMCQ-Gemma2-9b*	9	0.9108	0.8956	0.9031
	PMCQ-Llama3.1-8b	8	**0.9135**	**0.8964**	**0.9049**
	PMCQ-Mistral-7b	7	0.9117	0.8959	0.9037

Table 2: BERT Scores Between Generated Questions and Reference Questions

ity and effectiveness of the *PersianMCQ-Instruct* dataset for enhancing LLM performance. The next step in our study involved a detailed assessment of our models using the reserved evaluation texts. At the outset, we utilized well-known metrics like BERTScore. This metric enabled us to compare the quality of questions generated by our fine-tuned models against those from the *PersianMCQ-Instruct*. The outcomes, which are displayed in Table 2 , reveal that fine-tuning led to improved BERTScores across all models. However, it's noteworthy that BERTScore may not be the best indicator for assessing the quality of generated Persian MCQs due to certain inherent limitations. BERTScore has limitations in assessing Persian MCQs accurately. Some questions might still be relevant even if they don't match the *PersianMCQ-Instruct* dataset, and BERTScore overlooks grammatical and syntactic errors. Therefore, we bolstered our analysis with evaluations from human reviewers to ensure a more reliable assessment.

Additionally, we conducted a detailed human evaluation comparing the models' pre- and post-fine-tuning performances using the same five-level rating system described in section 3.1.4. We selected 200 questions from each model, both pre- and post-fine-tuning. Each expert evaluated 750 questions, with 300 questions in common, comprising 50 from each model. The agreement rate between the annotators was 0.96 and Cohen's kappa was 0.95. The comprehensive results of this assessment are thoroughly documented and can be found in Figure 5. As shown in Figure 5, all models exhibited improved performance after fine-tuning. In Appendix E we discuss the quality of the generated questions before and after fine-tuning and analyze instances where the model failed. Appendix B provides a range of examples, each accompanied by an explanation that outlines the reasoning behind its specific label.The evaluated data and code for the human annotation user interface (UI) are available on GitHub[20].

We also assigned values to the ratings as follows: A=5, B=4, C=3, D=2, and E=1, where a higher value indicates better performance. You can see the overall ratings in the table 3. After fine-tuning, overall performance across all models improved, demonstrating the quality of *PersianMCQ-Instruct*, as it was used as the dataset.

[20]https://github.com/KamyarZeinalipour/HumanAnnotation-UI-PMCQ

	Model	Ovr. Rating
Base	*gemma2-9b-it*	2.72
	Llama3.1-8b-Instruct	3.31
	Mistral-7b-Instruct-v0.3	1.30
Fine-tuned	*PMCQ-Gemma2-9b*	2.78
	PMCQ-Llama3.1-8b	4.51
	PMCQ-Mistral-7b	4.32
	PersianMCQ-Instruct	**4.75**

Table 3: Overall Ratings of the Models

5 Conclusion

In summary, this paper introduces the *PersianMCQ-Instruct* dataset, a comprehensive collection containing over 4000 unique texts and more than 1200 multiple-choice questions (MCQs) in Persian. This dataset provides both text content and corresponding MCQs in Persian. We rigorously evaluated the quality of this dataset through human assessment and automatic metrics, validating its reliability and effectiveness.
Moreover, we fine-tuned three different small-sized language models (LLMs) ranging from 7 billion to 9 billion parametersincluding *gemma2-9b-it*, *Llama3.1-8b-Instruct*, and *Mistral-7b-Instruct-v0.3*using this dataset. The resulting models, *PMCQ-Gemma2-9b*, *PMCQ-Llama3.1-8b*, and *PMCQ-Mistral-7b*, demonstrated a significant improvement in generating high-quality Persian MCQs. This underscores the dataset's utility and potential impact.
Our models and dataset are publicly available, paving the way for various educational applications in the Persian language. In this work, we also help tackle low-resource languages, improving the Persian language model. Looking ahead, we plan to expand the dataset with a greater focus on educational content across diverse subjects such as mathematics, physics, and history. Additionally, we aim to extend this initiative to other languages, broadening its applicability and impact.

6 Limitations

The *PersianMCQ-Instruct* resource has some limitations. The dataset is drawn exclusively from Persian Wikipedia, limiting topic diversity and question depth. While effective at generating fact-based questions, the models struggle with complex inference-based questions. Although human evaluations improved quality, a broader assessment with more Persian speakers would better gauge real-world utility. The models, due to low-resource language constraints, may miss subtle Persian nuances, and their large size requires substantial computational power, limiting accessibility. Additionally, their training on Wikipedia data restricts generalization to other educational topics, suggesting the need for further fine-tuning and dataset expansion.

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohamadali Nematbakhsh, and Arefeh Kazemi. 2021. Parsquad: Persian question answering dataset based on machine translation of squad 2.0. *International Journal of Web Research*, 4(1):34–46.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd workshop on machine reading for question answering*, pages 154–162.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2019. Natural question generation with reinforcement learning based graph-to-sequence model. *arXiv preprint arXiv:1910.08832*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2067–2073.

Liam Dugan, Eleni Miltsakaki, Shriyash Upadhyay, Etan Ginsberg, Hannah Gonzalez, Dayheon Choi, Chuning Yuan, and Chris Callison-Burch. 2022. A feasibility study of answer-agnostic question generation for education. *arXiv preprint arXiv:2203.08685*.

Yu-Shan Shih Gerald Rau. 2021. Evaluation of cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, 53.

Huanli Gong, Liangming Pan, and Hengchang Hu. 2022. Khanq: A dataset for generating deep questions in education. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5925–5938.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023a. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, 11:20885–20896.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Veronique Hoste, Chris Develder, and Thomas Demeester. 2023b. Diverse content selection for educational question generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 123–133.

Vrindavan Harrison and Marilyn Walker. 2018. Neural generation of diverse questions using answer focus, contextual and linguistic features. *arXiv preprint arXiv:1809.02637*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2021. Eqg-race: Examination-type question generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13143–13151.

Arefeh Kazemi, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. 2022. Persianquad: the native question answering dataset for the persian language. *IEEE Access*, 10:26045–26057.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, et al. 2021. Parsinlu: a suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Association for Computational Linguistics*, pages 74–81.

Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Xiyao Ma, Qile Zhu, Yanlin Zhou, and Xiaolin Li. 2020. Improving question generation with sentence-level semantic matching and answer position inferring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8464–8471.

Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codas, Yadong Lu, Wei-ge Chen, Olga Vrousgos, Corby Rosset, et al. 2024. Agentinstruct: Toward generative teaching with agentic flows. *arXiv preprint arXiv:2407.03502*.

Alireza Naeiji. 2022. Question generation using sequence-to-sequence model with semantic role labels.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests. *arXiv preprint arXiv:2011.13100*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *arXiv preprint arXiv:2401.06466*.

Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. *arXiv preprint arXiv:2106.04262*.

Tim Steuer, Anna Filighera, and Thomas Tregel. 2022. Investigating educational and noneducational answer selection for educational question generation. *IEEE Access*, 10:63522–63531.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3930–3939.

Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129.

Jiayun Wang, Wenge Rong, Jun Bai, Zhiwei Sun, Yuanxin Ouyang, and Zhang Xiong. 2023. Multi-source soft labeling and hard negative sampling for retrieval distractor ranking. *IEEE Transactions on Learning Technologies*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zichao Wang and Richard Baraniuk. 2023. Multiqg-ti: Towards question generation from multi-modal sources. *arXiv preprint arXiv:2307.04643*.

Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G Baraniuk. 2022b. Towards human-like educational question generation with large language models. In *International conference on artificial intelligence in education*, pages 153–166. Springer.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. *arXiv preprint arXiv:2203.13947*.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012*.

Kamyar Zeinalipour, Achille Fusco, Asya Zanollo, Marco Maggini, and Marco Gori. 2024a. Harnessing llms for educational content-driven italian crossword generation. *arXiv preprint arXiv:2411.16936*.

Kamyar Zeinalipour, Tommaso Iaquinta, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, and Marco Gori. 2023a. Building bridges of knowledge: Innovating education with automated crossword generation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1228–1236. IEEE.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024b. Automating turkish educational quiz generation using large language models. *arXiv preprint arXiv:2406.03397*.

Kamyar Zeinalipour, Yusuf Gökberk Kepti, Marco Maggini, Leonardo Rigutini, and Marco Gori. 2024c. A turkish educational crossword puzzle generator. *arXiv preprint arXiv:2405.07035*.

Kamyar Zeinalipour, Mohamed Zaky Saad, Marco Maggini, and Marco Gori. 2023b. Arabicros: Ai-powered arabic crossword puzzle generation for educational applications. *arXiv preprint arXiv:2312.01339*.

Kamyar Zeinalipour, Asya Zanollo, Giovanni Angelini, Leonardo Rigutini, Marco Maggini, and Marco Gori. 2023c. Italian crossword generator: Enhancing education through interactive word puzzles. *arXiv preprint arXiv:2311.15723*.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. *arXiv preprint arXiv:2203.14187*.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Multi-task learning with language modeling for question generation. *arXiv preprint arXiv:1908.11813*.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024. Clue-instruct: Text-based clue generation for educational crossword puzzles. *arXiv preprint arXiv:2404.06186*.

A Prompt Templates

Here, you can view all the various prompts we used in this study for the Persian MCQ task.

```
As an Educational Assistant your role is to transform the given Persian text into an
argumentative passage, supporting a specific viewpoint and main topic with relevant evidence.
Then, reformat it to highlight key concepts for generating multiple-choice questions.
1. Transform the original text into an argumentative passage.
2. Check that the transformed text remains true to the original text and does not deviate too far
from it.
3. Ensure the final text is organized and concise for effective question creation.
4. Proofread for grammar and spelling.
5. Ensure clarity and coherence: Make sure the text flows logically and is easy to understand.
6. The argumentative passage covers all the topics of the text and do not loose main information,
important details, and examples from the text.
The output should be a concise well-structured text in Persian.I just need the argumentative
passage as an output, nothing else.
\n\n TEXT: {original_text}
```

Figure 6: Text Augmentation Prompt

```
As an Educational Assistant, your role is to generate multiple-choice questions in Persian from the given text. Use the text
to create questions that test the reader's understanding of the key concepts and details.
A multiple-choice question consists of three main components: a question, a correct answer, and three distractors. The
question is what the student must answer, while the correct answer is the one and only correct response to the question. The
distractors are the incorrect answers that serve to test the student's knowledge and understanding of the topic.
1. Identify key concepts and details in the text.
2. Create clear, concise, and relevant multiple-choice questions and the question exists directly in the text.
3. Ensure each question has four answer choices, with one correct answer and three distractors.
4. Make sure the correct answer exists directly in the text and is true.
5. Make sure the question is grammatically correct in Persian and is easy for students to understand and should not be overly
complex or ambiguous.
6. The distractors should be plausible but incorrect answers that are related to the topic and the question. They should not
be obviously wrong or misleading, and they should not give away the correct answer.
7. Avoid using overly long sentences in the question, correct answer, or distractors.
8. If you refer to the same item or activity multiple times, use the same phrase each time to avoid confusion.
9. Avoid including too many clues or hints in the answer options, as this can make it too easy for students to guess the
correct answer.
10. Ensure that the questions are identical and not repetitive, even in content, to maintain variety and challenge.
11. Generate three questions from the text that maintain their quality, cover all the main topics, and are accurate.

Each string in the list should follow this format:
1. The first line contains the main question text.
2. The subsequent lines contain the answer options, each starting with one of the letters الف, ب, ج, or د, followed by a
period, the text of the option, and then an indicator in parentheses specifying whether it is the correct answer (پاسخ صحیح)
or an incorrect answer (پاسخ غلط).
3. Correct questions should always be the first choice in الف.

Here is an example with placeholders:


[
    "Question text 1?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد.
Incorrect answer (پاسخ غلط)",
    "Question text 2?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد.
Incorrect answer (پاسخ غلط)",
    "Question text 3?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ غلط)\nد.
Incorrect answer (پاسخ غلط)"
]

Do not include 'in' or 'sider' when generating questions; use them only for determining list members.

\n\nTEXT: {first_step_text}
```

Figure 7: Persian MCQ Generation Prompt

```
As an Educational Assistant, your role is to increase the complexity of the three multiple-choice questions that
you receive as input, using the text for this purpose.
1. For each question, make the structure of question more complex, requiring careful analysis and deeper thought
to understand it and it exists directly in  the text.
 2. Ensure that the correct answer and the three distractors are more nuanced and closely related, requiring
careful analysis to discern the correct option.
3. The correct answer and three distractors should be more complex in content and structure, and the correct
answer should be directly in the text and be true.
4. Incorporate common misconceptions or similar terminology to test deeper understanding.
5. Ensure that the distractors are plausible but falls  and subtle enough to challenge the reader's comprehension.
6. Make sure the questions and answers are  grammatically correct in Persian.
7. The questions and answers should not be  long when compared with the original questions.

Each string in the list should follow this format:
1. The first line contains the main question text.
2. The subsequent lines contain the answer options, each starting with one of the letters الف, ب, ج, or د,
followed by a period, the text of the option, and then an indicator in parentheses specifying whether it is the
correct answer (پاسخ صحیح) or an incorrect answer (پاسخ غلط).
3. Correct questions should always be the first choice in الف.
Here is an example with placeholders:


[
    "Question text 1?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ
غلط)\nد. Incorrect answer (پاسخ غلط)",
    "Question text 2?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ
غلط)\nد. Incorrect answer (پاسخ غلط)",
    "Question text 3?\nالف. Correct answer (پاسخ صحیح)\nب. Incorrect answer (پاسخ غلط)\nج. Incorrect answer (پاسخ
غلط)\nد. Incorrect answer (پاسخ غلط)"
]

Do not include 'in' or 'sider' when generating questions; use them only for determining list members.

\n\n TEXT: {first_step_text} \n Multiple-Choice Questions: {second_step_text}
```

Figure 8: Persian MCQ Generation Prompt

B Example of Generated Perisan MCQs with LLMs on *PersianMCQ-Instruct*

Table 4 presents a comparison of generated questions and answers produced before and after fine-tuning the models using *PersianMCQ-Instruct* and also sample ratings labeled A, B, C and D, and, Table 5 includes along with their translations. Since there are no samples for rating E in the generated questions and answers in Table 4, additional examples were extracted and are shown in Tables 6 7 8 9. These tables include both the original Persian text and English translations to facilitate comprehension for non-Persian speakers.

گلنکور (به انگلیسی: erocnelG) شرکت بریتانیایی-سوئیسی چندملیتی صنایع معدنی و تجارت کالا می‌باشد، که دفتر مرکزی آن در شهر بر، سوئیس قرار دارد و دفتر عملیاتی آن در سن هلیه، جرزی، بریتانیا مستقر می‌باشد. شرکت گلنکور بزرگترین شرکت کالاهای اقتصادی در جهان به‌شمار می‌آید، همچنین در مجموع از نظر میزان درآمد سالیانه، به‌عنوان بزرگترین شرکت سوئیسی نیز محسوب می‌شود. این شرکت در سال دوهزارو سیزده در فهرست فورچون جهانی پانصد در رتبه دوازده از بزرگترین شرکت‌های جهان قرار گرفت. شرکت گلنکور در سال دو هزار و سیزده شرکت اکستراتا، که یکی از بزرگترین شرکت‌های معدنی جهان است را خریداری نمود، سپس اکستراتا منحل شده و دارایی‌های آن در گلنکور ادغام گردید. هم‌اکنون گلنکور به‌عنوان یکی از بزرگترین شرکت‌های استخراج معادن جهان شناخته می‌شود.			
gemma2-9b-it	دفتر مرکزی شرکت گلنکور در کدام شهر واقع شده است؟	الف. سن هلیه (پاسخ صحیح) ب. لندن ج. برلین د. ژنو	C جواب غلط است و در تعیین جواب درست عدم توانایی داشته .
Continued on next page			

گلنکور (به انگلیسی: erocnelG) شرکت بریتانیایی-سوئیسی چندملیتی صنایع معدنی و تجارت کالا می‌باشد، که دفتر مرکزی آن در شهر بر، سوئیس قرار دارد و دفتر عملیاتی آن در سن هلیه، جرزی، بریتانیا مستقر می‌باشد. شرکت گلنکور بزرگترین شرکت کالاهای اقتصادی در جهان به‌شمار می‌آید، همچنین در مجموع از نظر میزان درآمد سالیانه، به‌عنوان بزرگترین شرکت سوئیسی نیز محسوب می‌شود. این شرکت در سال دوهزارو سیزده در فهرست فورچون جهانی پانصد در رتبه دوازده از بزرگترین شرکت‌های جهان قرار گرفت. شرکت گلنکور در سال دو هزار و سیزده شرکت اکستراتا، که یکی از بزرگترین شرکت‌های معدنی جهان است را خریداری نمود، سپس اکستراتا منحل شده و دارایی‌های آن در گلنکور ادغام گردید. هم‌اکنون گلنکور به‌عنوان یکی از بزرگترین شرکت‌های استخراج معادن جهان شناخته می‌شود.			
PMCQ-Gemma2-9b	گلنکور به عنوان یک شرکت بین‌المللی فعال در کدام دو کشور واقع شده‌است؟	الف. بریتانیای کبیر و سوئیس (پاسخ صحیح) ب. امریکا و کانادا (پاسخ غلط) ج. استرالیا و ژاپن (پاسخ غلط) د. فرانسه و آلمان (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.
Llama3.1-8b-Instruct	محل دفتر مرکزی شرکت گلنکور کجاست؟	الف. شهر لندن، انگلستان ب. شهر بر، سوئیس ج. شهر پاریس، فرانسه د. شهر نیویورک، آمریکا	B عدم تعیین پاسخ‌های صحیح و غلط.
PMCQ-Llama3.1-8b	دفتر مرکزی شرکت گلنکور که نقش کلیدی در مدیریت فعالیت‌های بین‌المللی آن ایفا می‌کند، در کدام کشور واقع شده است؟	الف. سوئیس (پاسخ صحیح) ب. انگلستان (پاسخ غلط) ج. کانادا (پاسخ غلط) د. استرالیا (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.
Mistral-7b-Instruct-v0.3	شرکت گلنکور در چه سالی تاسس شد؟	الف- ۱۳۷۲ ب-۲۳۷۱ ج-۱۴۰۰ د- ۱۳۸۰	D عدم وجود سوال در متن.
PMCQ-Mistral-7b	دفتر مرکزی شرکت گلنکور که به عنوان یکی از بزرگترین شرکت‌های جهان شناخته می‌شود، در کدام شهر قرار دارد؟	الف. بر، سوئیس (پاسخ صحیح) ب. سن هلیه، جرزی (پاسخ غلط) ج. لندن، انگلستان (پاسخ غلط) د. پاریس، فرانسه (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.

Table 4: Comparison of Results Across Each Model Before and After Fine-Tuning

Glencore is a British-Swiss multinational company in the mining and commodity trading industries, headquartered in Baar, Switzerland, with its operational office located in St. Helier, Jersey, United Kingdom. Glencore is considered the largest commodity trading company in the world and, in terms of annual revenue, is also regarded as the largest Swiss company. In 2013, it ranked 12th on the Fortune Global 500 list of the world's largest companies. In the same year, Glencore acquired Xstrata, one of the largest mining companies in the world. Following the acquisition, Xstrata was dissolved, and its assets were merged into Glencore. Today, Glencore is recognized as one of the largest mining companies globally.			
Model	Question	Answer	Rate
Continued on next page			

Glencore is a British-Swiss multinational company in the mining and commodity trading industries, headquartered in Baar, Switzerland, with its operational office located in St. Helier, Jersey, United Kingdom. Glencore is considered the largest commodity trading company in the world and, in terms of annual revenue, is also regarded as the largest Swiss company. In 2013, it ranked 12th on the Fortune Global 500 list of the world's largest companies. In the same year, Glencore acquired Xstrata, one of the largest mining companies in the world. Following the acquisition, Xstrata was dissolved, and its assets were merged into Glencore. Today, Glencore is recognized as one of the largest mining companies globally.			
gemma2-9b-it	Where is the headquarters of Glencore located?	a. St. Helier (Correct Answer) b. London c. Berlin d. Geneva	C Incorrect answer and failure to identify the incorrect option in wrong answers.
PMCQ-Gemma2-9b	In which two countries is Glencore, as an international company, located?	a. Great Britain and Switzerland (Correct Answer) b. USA and Canada (Incorrect) c. Australia and Japan (Incorrect) d. France and Germany (Incorrect)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
PMCQ-Llama3.1-8b	Where is the headquarters of Glencore located?	a. London, England b. Baar, Switzerland (Correct Answer) c. Paris, France d. New York, USA	B Failure to distinguish between correct and incorrect answers.
Llama3.1-8b-Instruct	The headquarters of Glencore, which plays a key role in managing its international activities, is located in which country?	a. Switzerland (Correct Answer) b. United Kingdom (Incorrect) c. Canada (Incorrect) d. Australia (Incorrect)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
Mistral-7b-Instruct-v0.3	In what year was Glencore company founded?	a. 1372 b. 2371 c. 1400 d. 1380	D Question contains factual inaccuracies
PMCQ-Mistral-7b	As one of the largest mining companies in the world, where is the headquarters of Glencore located?	a. Baar, Switzerland (Correct Answer) b. St. Helier, Jersey (Incorrect) c. London, England (Incorrect) d. Paris, France (Incorrect)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.

Table 5: Translation of Comparison of Results Across Each Model Before and After Fine-Tuning

پس از میلاد یا آنُو دُمینی (لاتین: inimoD onna) که مخفف آن "DA" است معرف یک مبدأ تاریخ است که بر پایه سال باستانی - تخمینی مورد قبول تولد عیسی مسیح قرار دارد. به همین شکل، «پیش از میلاد» (از یونان باستان « کریستوس» یا تدهین شده «اشاره به عیسی مسیح»)، مخفف شده به «بی‌سی» "CB" که در زبان انگلیسی مورد استفاده قرار گرفته و به سال‌های قبل از شروع این مبدأ اشاره دارد. برخی از افراد غیر-مسیحی از مخفف‌های ای‌دی و بی‌سی بدون اشاره به مسیحی بودن آن، استفاده می‌کنند. بعضی از مردم ترجیح می‌دهند از عبارات جایگزین مانند «سی ای» و «بی‌سی‌ای» استفاده نمایند و علت آن را نیز خنثی بودن بیشتر این نوع الفاظ ذکر می‌کنند (رجوع کنید به‌عصر مشترک).
این تخصیص جهت شمارش سالیان در «عصر مسیحیت» و به صورت مرسوم در ژولیوس و تقویم مسیحی به کار گرفته می‌شود. به صورت کامل‌تر، شمار سال‌ها را همچنین می‌توان به عنوان «انو دامینی نوستری ایسو کریستی» («در سال‌های سرور ما عیسی مسیح»). تاریخ گذاری «انو دامینی» برای اولین بار در سال ۵۲۵ مورد محاسبه قرار گرفته و در سال‌های قرن هجدهم در اروپای غربی استفاده از آن شروع شد.
شمار سالیان بر اساس عصر مسیحیت اکنون در بسیاری از مناطق جهان مورد استفاده قرار دارد. این استفاده هم شامل موارد تجاری و هم موارد علمی می گردد. برای دهه‌های متوالی، این تاریخ به عنوان استاندارد جهانی پذیرفته‌شده‌است و توسط سازمان‌های بین‌المللی مانند سازمان ملل و اتحادیه پستی جهانی مورد استفاده قرار گرفته‌است. این موضوع بخاطر رواج مسیحیت در دنیای غرب، تأثیر فوق‌العاده دنیای غرب بر علوم، فناوری و تجارت و همچنین این واقعیت که تاریخ خورشیدی تقویم مسیحیت برای زمانی دراز، به عنوان ابزار صحت فضائی در نظر گرفته می‌شود، است.
انگلیسی به کپی کردن استفاده لاتین در قرار دادن حروف مخفف «قبل» از رقم سال برای ای دی پرداخته ولی بی سی را «بعد» از رقم سال قرار می‌دهد؛ به عنوان مثال ۴۶ بی سی، ولی ای دی ۶۰۰۲.

Model	Question	Answer	Rate
Mistral-7b-Instruct-v0.3		الف. تاریخ گذاری "انو دامینی" برای اولین بار در سال ۵۲۵ مورد محاسبه قرار گرفته است. (rewsnA tcerroC) ب. تاریخ گذاری "انو دامینی" برای اولین بار در سال ۵۰۰۲ مورد محاسبه قرار گرفته است. (rewsnA tcerrocnI) ج. تاریخ گذاری "انو دامینی" برای اولین بار در سال ۵۰۰۱ مورد محاسبه قرار گرفته است. (rewsnA tcerrocnI) د. تاریخ گذاری "انو دامینی" برای اولین بار در سال ۵۰۰۳ مورد محاسبه قرار گرفته است. (rewsnA tcerrocnI)	E سوال تولید نشده است.

Table 6: Sample of rating E

<table>
<tr><td colspan="4">After Christ, or Anno Domini (Latin: anno Domini), abbreviated as "AD," denotes a historical era based on the traditionally accepted approximate birth year of Jesus Christ. Similarly, "Before Christ" (derived from Ancient Greek Christos, meaning "the Anointed One," referring to Jesus Christ), abbreviated as "BC," is used in English to refer to years before the beginning of this era. Some non-Christians use the abbreviations AD and BC without religious connotation. Others prefer alternative terms, such as "CE" (Common Era) and "BCE" (Before Common Era), considering these expressions to be more neutral (see Common Era).
This system of year counting in the "Christian Era" is conventionally used within the Julian and Christian calendars. More precisely, the year count can also be referred to as anno Domini nostri Iesu Christi ("in the year of our Lord Jesus Christ"). Anno Domini dating was first calculated in the year 525 and began to be used in Western Europe in the 18th century.
The system of dating based on the Christian era is now widely used across various regions of the world, for both commercial and scientific purposes. For decades, this calendar has been recognized as a global standard and is used by international organizations, such as the United Nations and the Universal Postal Union. This adoption is due to the widespread influence of Christianity in the Western world, the tremendous impact of the West on science, technology, and commerce, and the fact that the solar dating of the Christian calendar has long been considered a reliable tool for spatial accuracy.
English borrows the Latin usage of placing the abbreviation "AD" before the year number, while "BC" is placed after the year number. For example, 64 BC but AD 2006.</td></tr>
<tr><td>Model</td><td>Question</td><td>Answer</td><td>Rate</td></tr>
<tr><td>Mistral-7b-Instruct-v0.3</td><td></td><td>a. The "Anno Domini" dating was first calculated in the year 525. (Correct Answer) b. The "Anno Domini" dating was first calculated in the year 2005. (Incorrect Answer) c. The "Anno Domini" dating was first calculated in the year 1005. (Incorrect Answer) d. The "Anno Domini" dating was first calculated in the year 3005. (Incorrect Answer)</td><td>E
Question is Empty</td></tr>
</table>

Table 7: Translation of sample for rating E

هواپیما (به انگلیسی: enalpria یا enalporea) (حالت غیررسمی: enalp)، هواگرد ثابت‌بالی است که توسط موتور جت، ملخ یا موتور راکت به جلو رانده می‌شود. هواپیماها در ابعاد، اشکال و آرایش‌های مختلف بال ظاهر می‌شوند. طیف وسیع کاربردهای هواپیما شامل مواردی از قبیل تفریح، انتقال کالا و افراد، نظامی و تحقیقاتی است. انتقال‌های هوانوردی تجاری در سراسر جهان سالانه بیش از چهار میلیارد مسافر را از طریق هواپیماهای مسافربری و بیش از ۰۰۲ میلیارد تن-کیلومتر محموله را سالانه جابجا کرده که کمتر از ۱ درصد از جابجایی محموله‌ها در سراسر جهان است. اکثر هواپیماها توسط خلبانی از اعضای خدمه به پرواز در می‌آید، اما برخی از هواپیماها همچون پهپادها به گونه‌ای طراحی شده‌اند که توسط کنترل از راه دور یا رایانه به پرواز درآیند. برادران رایت اولین هواپیما را در ۳۰۹۱ میلادی اختراع کرده و به پرواز درآوردند و اختراعشان به عنوان «اولین پرنده سنگین‌تر از هوای پایدار و کنترل‌شده» شناخته شد. کار آن‌ها ادامه کارهای جورج کیلی در ۹۹۷۱ بود، کسی که مفهوم هواپیماهای نوین را به پیش کشید (سپس مدل‌ها و گلایدرهای حامل مسافر را ساخته و با موفقیت به پرواز درآورد). اوتو لیلینتال بین سال‌های ۷۶۸۱ تا ۶۹۸۱ میلادی، جزو پیشتازان هوانوردی انسانی اهل آلمان بود که به مطالعه پرندگان سنگین‌تر-از-هوا نیز پرداخت. فناوری هواپیماها پس از استفاده محدود ازشان در جنگ جهانی اول، به توسعه خود ادامه داد. هواپیماها در تمامی نبردهای عمده جنگ جهانی دوم حضور داشته‌اند. اولین هواپیمای جت، هاینکل ۸۷۱ آلمانی در ۹۳۹۱ میلادی بود. اولین جت مسافربری، د هویلند کامت بود که در ۲۵۹۱ میلادی معرفی شد. بوئینگ ۷۰۷، اولین جت تجاری بود که به موفقیت گسترده دست یافت و تا بیش از ۰۵ سال از ۸۵۹۱ تا ۳۱۰۲ میلادی مشغول خدمت بود.			
Model	Question	Answer	Rate
gemma2-9b-it	الف. کسانی که اولین هواپیما را اختراع کرده و به پرواز درآوردند؟	ب. جورج کیلی ج. برادران رایت د. اوتو لیلینتال	E سوال تولید نشده و به عنوان یکی از گزینه ها در نظر گرفته شده است.

Table 8: Sample of rating E

An airplane (informally: plane) is a fixed-wing aircraft propelled forward by jet engines, propellers, or rocket engines. Airplanes come in various sizes, shapes, and wing configurations and serve a wide range of purposes, including leisure, cargo and passenger transportation, military use, and research. Globally, commercial aviation transports over four billion passengers annually on passenger planes and moves more than 200 billion ton-kilometers of cargo each year, which accounts for less than 1% of worldwide cargo transport. Most airplanes are piloted by crew members, but some, like drones, are designed to be remotely controlled or operated by computers. The Wright brothers invented and flew the first airplane in 1903, recognized as the "first stable, controlled heavier-than-air flight." Their work built on the concepts established by George Cayley in 1799, who pioneered modern airplane ideas and later built and successfully flew passenger-carrying models and gliders. Otto Lilienthal, a German aviation pioneer from 1867 to 1896, also contributed to the study of heavier-than-air flight. Airplane technology continued to develop after limited use in World War I, and airplanes played significant roles in all major battles of World War II. The first jet airplane was the German Heinkel 178 in 1939, and the first commercial jet airliner was the de Havilland Comet, introduced in 1952. Boeings 707 was the first widely successful commercial jet, serving for over 50 years, from 1958 until 2013.			
Model	Question	Answer	Rate
gemma2-9b-it	A. the people who invented and flew the first airplane?	B. George Cayley C. The Wright Brothers D. Otto Lilienthal	E The question has not been generated and is considered as one of the options.

Table 9: Translation of sample for rating E

C Example of *PersianMCQ-Instruct*

Table 10 provides a comparison of generated questions and answers, as well as more complex examples of *PersianMCQ-Instruct* , along with sample ratings labeled A and C. Table 11 includes these examples alongside their translations. Since there are no examples with a rating of B in Table 10, additional samples are shown in Tables 12 and 13. These tables present both the original Persian text and English translations to assist non-Persian speakers. Notably, *PersianMCQ-Instruct* did not contain any questions or answers with a rating of E.

<table>
<tr><td colspan="3">حَواریون (جمع حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود.حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود. تمام ۲۱ حواری در ابتدا تابع دین یهودیت و همگی (همچون خود عیسی‌مسیح) اصالتاً یهودی‌زاده (یهودی‌نژاد) بودند. ایشان پس از ایمان به عیسی‌مسیح، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که از نژاد یهود بودند، اما به عیسی‌مسیح ایمان آورده بودند و به این ترتیب دیانت ایشان از یهودیت به مسیحیت تغییر می‌یافت. علت تأکید بر این موضوع این است که "یهودی" بودن را می‌توان از دو دیدگاه تعریف کرد: «نژاد» و «دیانت». حَواریون، به معنی یاران برگزیده، عنوان دوازده تن از شاگردان و یاران ویژه عیسی مسیح است. واژه "اپاسل" از زبان یونانی به معنای "رسولان" گرفته شده و به خاطر این شباهت، اصطلاح "رسولان" نیز برای حواریون استفاده می‌شود. این حواریون در ابتدا از دین یهودیت پیروی می کردند و همگی از نژاد یهودی بودند. دلیل برگزیده شدن این حواریون به عنوان شاگردان ویژه عیسی مسیح، تغییر ایمان آن‌ها از یهودیت به مسیحیت بود. آنان پس از ایمان به عیسی، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که همچنان از نژاد یهود بودند اما ایمان‌های دینی‌شان به مسیحیت تغییر یافته بود. این موضوع اهمیت دارد زیرا یهودیت را می‌توان از دو دیدگاه بررسی کرد: یکی نژاد و دیگری دیانت. تغییر دیانت این افراد از یهودیت به مسیحیت نشان‌دهنده منزوی کردن نژاد از دیانت است و به اهمیت تحول در ایمان و شناخت معنوی آن‌ها اشاره دارد. این تغییر نه تنها بر افزایش شمار پیروان مسیحیت تأثیرگذار بود، بلکه نمادی از انتقال معنوی و فرهنگ دینی بود.</td></tr>
<tr><td>Question</td><td>Answer</td><td>Rate</td></tr>
<tr><td>حواریون به چه معناست؟</td><td>الف. یاران برگزیده (پاسخ صحیح) ب. پیامبران (پاسخ غلط) ج. رهبران مذهبی (پاسخ غلط) د. محققان دینی (پاسخ غلط)</td><td>A
تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.</td></tr>
<tr><td>واژه یونانی 'اپاسل' به چه معناست؟</td><td>الف. رسولان (پاسخ صحیح) ب. معلمان (پاسخ غلط) ج. یاران (پاسخ غلط) د. پیامبران (پاسخ غلط)</td><td>A
تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.</td></tr>
</table>

Continued on next page

حَواریون (جمع حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود.حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود. تمام ۲۱ حواری در ابتدا تابع دین یهودیت و همگی (همچون خود عیسی‌مسیح) اصالتاً یهودی‌زاده (یهودی‌نژاد) بودند. ایشان پس از ایمان به عیسی‌مسیح، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که از نژاد یهود بودند، اما به عیسی‌مسیح ایمان آورده بودند و به این ترتیب دیانت ایشان از یهودیت به مسیحیت تغییر می‌یافت. علت تأکید بر این موضوع این است که "یهودی" بودن را می‌توان از دو دیدگاه تعریف کرد: «نژاد» و «دیانت». حَواریون، به معنی یاران برگزیده، عنوان دوازده تن از شاگردان و یاران ویژه عیسی مسیح است. واژه "اپاسل" از زبان یونانی به معنای "رسولان" گرفته شده و به خاطر این شباهت، اصطلاح "رسولان" نیز برای حواریون استفاده می‌شود. این حواریون در ابتدا از دین یهودیت پیروی می کردند و همگی از نژاد یهودی بودند. دلیل برگزیده شدن این حواریون به عنوان شاگردان ویژه عیسی مسیح، تغییر ایمان آن‌ها از یهودیت به مسیحیت بود. آنان پس از ایمان به عیسی، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که همچنان از نژاد یهود بودند اما ایمان‌های دینی‌شان به مسیحیت تغییر یافته بود. این موضوع اهمیت دارد زیرا یهودیت را می‌توان از دو دیدگاه بررسی کرد: یکی نژاد و دیگری دیانت. تغییر دیانت این افراد از یهودیت به مسیحیت نشان‌دهنده منزوی کردن نژاد از دیانت است و به اهمیت تحول در ایمان و شناخت معنوی آن‌ها اشاره دارد. این تغییر نه تنها بر افزایش شمار پیروان مسیحیت تأثیرگذار بود، بلکه نمادی از انتقال معنوی و فرهنگ دینی بود.

Question	Answer	Rate
تغییر دیانت حواریون از یهودیت به مسیحیت نشان‌دهنده چیست؟	الف. اهمیت تحول در ایمان و شناخت معنوی آنان (پاسخ صحیح) ب. عدم تغییر در روند تاریخی (پاسخ غلط) ج. افزایش دشمنی میان ادیان (پاسخ غلط) د. کاهش اهمیت دیانت (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.
More complex Question	Answer	Rate
حواریون در ابتدا پیرو کدام دیانت بودند و تغییر ایمانشان به چه معنا بود؟	الف. یهودیت؛ اهمیت تحول در ایمان و شناخت معنوی آنان (پاسخ صحیح) ب. مسیحیت؛ نشان‌دهنده انزوای نژادی آنان (پاسخ غلط) ج. اسلام؛ تأثیرگذار بر افزایش شمار پیروان مسیحیت (پاسخ غلط) د. زرتشتی؛ نماد انتقال فرهنگی و دینی (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاً صحیح.

Continued on next page

حَواریون (جمع حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود.حَواریون (جمع حَواری به معنی یارِ برگزیده) (به انگلیسی: eltsopA) عنوان دوازده تن از یاران و شاگردان ویژهٔ عیسی مسیح است. واژهٔ اپاسل (eltsopA) از آپوستولوی یونانی گرفته شده که به معنی «رسولان» می‌باشد و به همین علت از اصطلاح «رسولان» نیز برای حواریون استفاده می‌شود. تمام ۲۱ حواری در ابتدا تابع دین یهودیت و همگی (همچون خود عیسی‌مسیح) اصالتاً یهودی‌زاده (یهودی‌نژاد) بودند. ایشان پس از ایمان به عیسی‌مسیح، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که از نژاد یهود بودند، اما به عیسی‌مسیح ایمان آورده بودند و به این ترتیب دیانت ایشان از یهودیت به مسیحیت تغییر می‌یافت. علت تأکید بر این موضوع این است که "یهودی" بودن را می‌توان از دو دیدگاه تعریف کرد: «نژاد» و «دیانت»). حَواریون، به معنی یاران برگزیده، عنوان دوازده تن از شاگردان و یاران ویژه عیسی مسیح است. واژه "اپاسل" از زبان یونانی به معنای "رسولان" گرفته شده و به خاطر این شباهت، اصطلاح "رسولان" نیز برای حواریون استفاده می‌شود. این حواریون در ابتدا از دین یهودیت پیروی می کردند و همگی از نژاد یهودی بودند. دلیل برگزیده شدن این حواریون به عنوان شاگردان ویژه عیسی مسیح، تغییر ایمان آن‌ها از یهودیت به مسیحیت بود. آنان پس از ایمان به عیسی، جزو مسیحیان یهودی‌نژاد به شمار می‌آمدند، چرا که همچنان از نژاد یهود بودند اما ایمان‌های دینی‌شان به مسیحیت تغییر یافته بود. این موضوع اهمیت دارد زیرا یهودیت را می‌توان از دو دیدگاه بررسی کرد: یکی نژاد و دیگری دیانت. تغییر دیانت این افراد از یهودیت به مسیحیت نشان‌دهنده منزوی کردن نژاد از دیانت است و به اهمیت تحول در ایمان و شناخت معنوی آن‌ها اشاره دارد. این تغییر نه تنها بر افزایش شمار پیروان مسیحیت تأثیرگذار بود، بلکه نمادی از انتقال معنوی و فرهنگ دینی بود.		
Question	Answer	Rate
واژه یونانی 'اپاسل' به چه اصطلاحی و چرا اُستفاده می‌شود؟	الف. رسولان؛ به خاطر شباهت معنایی با حواریون (پاسخ صحیح) ب. محبان؛ به دلیل روح بزرگی آنان (پاسخ غلط) ج. شوراها؛ بخاطر اعتبار دینی (پاسخ غلط) د. مفسران، به دلیل نقش تفسیری (پاسخ غلط)	C سوال دارای غلط گرامری جدی میباشد.
چگونه می‌توان اهمیت تغییر دیانت حواریون را مورد ارزیابی قرار داد؟	الف. به عنوان نمادی از تحول معنوی و فرهنگی (پاسخ صحیح) ب. به واسطه افزایش دشمنی‌های دینی (پاسخ غلط) ج. با توجه به ثبات پیروان یهودیت (پاسخ غلط) د. بر اساس انسجام نژادی آنها (پاسخ غلط)	A تولید سوال و جواب کاملاً درست و مرتبط با متن با دستور زبان کاملاَ صحیح.

Table 10: Comparison of Questions Generated by GPT-4o (PersianMCQ-Instruct) and Sample Ratings A and C

The Apostles (plural of Apostle, meaning "chosen companion") refer to the twelve special disciples and companions of Jesus Christ. The word "Apostle" comes from the Greek term *apostoloi*, meaning "messengers," and thus the term "messengers" is also used to describe the Apostles. Initially, all twelve Apostles followed Judaism, and, like Jesus himself, they were of Jewish origin (Jewish by race). After they placed their faith in Jesus Christ, they were regarded as Jewish Christians, as they were ethnically Jewish but had converted to Christianity. The emphasis on this topic stems from the fact that "Jewish" can be defined from two perspectives: "ethnicity" and "religion." The change in their faith from Judaism to Christianity is essential because it highlights the separation between race and religion, pointing to the importance of their spiritual transformation and newfound understanding. This shift not only contributed to the growth in Christian followers but also symbolized a spiritual and cultural shift. The Apostles' decision to follow Jesus and change their beliefs represented a significant step in the spread of Christianity and the emergence of a new religious identity.		
Question	Answer	Rate
What does apostles mean?	A. Chosen companions (correct answer) B. Prophets (incorrect answer) C. Religious leaders (incorrect answer) D. Religious scholars (incorrect answer)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
What does the Aegean Greek 'Apasal' mean?	A. Apostles (correct answer) B. Teachers (incorrect answer) C. Companions (incorrect answer) D. Prophets (incorrect answer)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
What does the conversion of the apostles from Judaism to Christianity indicate?	A. The importance of the transformation in their faith and spiritual understanding (correct answer) B. No change in the historical process (incorrect answer) C. Increased hostility between religions (incorrect answer) D. Decreased importance of religion (incorrect answer)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
Continued on next page		

More Complex Question	Answer	Rate
Which religion did the apostles follow in the beginning and what was the meaning of their change of faith?	A. Judaism; the importance of the transformation in their faith and spiritual understanding (correct answer) B. Christianity; indicating their ethnic isolation (incorrect answer) C. Islam; influential in increasing the number of Christian followers (incorrect answer) D. Zoroastrianism; a symbol of cultural and religious transfer (incorrect answer)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.
What phrase the Greek word 'apostle' , and why is it osed?	A. Apostles; due to their semantic similarity with the disciples (correct answer) B. Lovers; because of their great spirit (incorrect answer) C. Councils; due t their religious authority (incorrect answer) D. Commentators; because of their interpretative role (incorrect answer)	C The question has serious grammatical issue.
How can the importance of the apostles' conversion be evaluated?	A. As a symbol of spiritual and cultural transformation (correct answer) B. Due to increased religious hostilities (incorrect answer) C. Considering the stability of Jewish followers (incorrect answer) D. Based on their ethnic cohesion (incorrect answer)	A Generation of completely correct and contextually relevant questions and answers with entirely accurate grammar.

Table 12: Translation of comparison of Questions Generated by GPT-4o (PersianMCQ-Instruct) and Sample Ratings A and C (continued)

<table>
<tr><td colspan="3">آشوری‌ها، مردمی هستند که ریشه آن‌ها به مردم سامی در بین‌النهرین باستان می‌رسد. آن‌ها علاوه بر زبان محل سکونت خود، به زبان آرامی نو آشوری (زبانی از خانواده زبان‌های سامی) سخن می گویند و بیشتر آن‌ها پیرو مسیحیت سریانی هستند. سرزمین آشوری‌ها اکنون در شمال کشور عراق (استان‌های دهوک و نینوا)، جنوب شرقی ترکیه (در منطقه‌ی حکاری و طور عبدین)، شمال شرقی سوریه (استان حسکه) قرار دارد. در طول قرن گذشته میلادی بسیاری از آشوری‌ها به نقاط دیگر دنیا از جمله قفقاز، آمریکای شمالی، اروپا و استرالیا مهاجرت کردند. حوادثی چون کشتارهای دیاربکر، نسل کشی آشوری‌ها (همراه با نسل کشی یونانی‌ها و نسل کشی ارمنی‌ها) توسط امپراتوری عثمانی در طول جنگ جهانی اول، کشتار سمیل در ۳۳۹۱ در عراق، سیاست‌های ملی گرایی عربی، حزب بعث عراق و انقلاب ۷۵۳۱ ایران، حمله داعش و اشغال مناطقی از عراق و سوریه، از عوامل برون کوچی آشوری‌ها بوده‌اند. تخمین زده می‌شود که تعداد آشوری‌های جهان ۵ میلیون نفر باشد. تعداد آشوری‌های ایران تا پیش از انقلاب ۷۵۳۱ حدود ۰۷ تا ۰۹ هزار نفر بود. بیشتر جمعیت آشوری‌ها در شهرهای ارومیه، تهران، شاهین شهر، سلماس، تبریز، اهواز، همدان، کرمانشاه، شیراز، مشهد، بندرعباس، ماهشهر، بابلسر، بندر انزلی، فردیس و قزوین واقع شده‌است. آشوری‌ها در قرون اول اسلامی با ترجمهٔ علوم و معارف دنیای باستان به زبان عربی خدمت بزرگی به توسعهٔ دانش در میان مسلمانان و نگهداری آثار علمی دوران باستان کردند. بعضی از بزرگ‌ترین دانشمندان و مترجمان خلافت اسلامی از این مردم بودند. آشوری‌ها بعد از قرون وسطی مورد توجه کلیسای کاتولیک و پس از آن کلیسای پروتستان قرار گرفتند. در جریان حوادث ناشی از جنگ جهانی اول و حضور ارتش‌های روسیه و عثمانی در آذربایجان و جنگ‌های قومی و مذهبی که پس از کشتن مارشیمون بنیامین، رهبر آشوری‌ها، آسیب‌های فراوان دیدند و بسیاری از آن‌ها ناگزیر به نقاط دیگر مهاجرت کردند. آشوری‌ها، مردمی با ریشه‌های سامی در بین‌النهرین باستان، نقش و اهمیت فوق‌العاده‌ای در حفظ و گسترش دانش و فرهنگ در طول تاریخ داشته‌اند. این قوم که به زبان آرامی نو آشوری سخن می گویند، عمدتاً پیرو مسیحیت سریانی هستند و در مناطقی از عراق، ترکیه، و سوریه سکونت دارند. اما در قرن گذشته، حوادث تراژیک متعدد آنان را وادار به مهاجرت به کشورهای دیگری مانند قفقاز، آمریکای شمالی، اروپا، و استرالیا کرده است. کشتارهای دیاربکر، نسل کشی آشوری‌ها به دست امپراتوری عثمانی، کشتار سمیل، سیاست‌های ملی گرایانه، و حملات داعش از جمله این عوامل بوده‌اند. تخمین زده می‌شود که جمعیت آشوری‌ها در جهان حدود ۵ میلیون نفر باشد، با جمعیتی قابل توجه در ایران در شهرهایی مانند ارومیه، تهران، و تبریز. این قوم از لحاظ دینی به کلیسای کاتولیک کلدانی، کلیسای آشوری مشرق، و کلیسای باستانی مشرق تعلق دارند و نقش مهمی در توسعه مسیحیت و علوم در دوره‌های مختلف ایفا کرده‌اند. آشوری‌ها، به ویژه در قرون اول اسلامی با ترجمه علوم و معارف به زبان عربی، خدمت بزرگی به توسعه دانش در میان مسلمانان کردند. در نهایت، با در نظر گرفتن مشکلات و مهاجرت‌های مکرر، آشوری‌ها همچنان پایداری خود را در حفظ فرهنگ و هویت خود نشان داده‌اند. این قوم همچنان بازمانده‌ای از قوم باستانی آشور، نواده نوح پیامبر، هستند و اهمیت تاریخی و فرهنگی خود را حتی در مواجهه با چالش‌ها حفظ کرده‌اند.</td></tr>
<tr><td>Question</td><td>Answer</td><td>Rate</td></tr>
<tr><td>آشوری‌ها به‌طور عمده به کدام شاخه‌های مسیحیت تعلق دارند و همچنین چه نقشی در توسعه مسیحیت ایفا کرده‌اند؟</td><td>الف. کلیسای کاتولیک کلدانی، کلیسای آشوری مشرق، و کلیسای باستانی مشرق؛ نقش مهم در توسعه علوم مسیحی و اسلامی (پاسخ صحیح) ب. کلیسای ارتدکس یونانی، کلیسای روسی، و کلیسای قبطی؛ نقش جزئی در ترجمه متون دینی (پاسخ غلط) ج. کلیسای پروتستان آلمان، کلیسای انجیلی آمریکا، و کلیسای کاتولیک روم؛ هیچ نقش قابل ملاحظه‌ای (پاسخ غلط) د. کلیسای روسی، کلیسای ارمنی، و کلیسای اوکراینی؛ تأثیرگذاری زیاد در ادبیات مسیحی (پاسخ غلط)</td><td>B
قسمت دوم جواب دارای اشتباه می‌باشد.</td></tr>
</table>

Table 13: Sample of rating B for GPT-4o

<table>
<tr><td colspan="3">The Assyrians are a people whose roots trace back to the Semitic people of ancient Mesopotamia. They speak not only the local language of their residence but also Modern Assyrian Aramaic (a language from the Semitic language family), and most of them are followers of Syriac Christianity. The land of the Assyrians is now located in northern Iraq (the Duhok and Nineveh provinces), southeastern Turkey (in the Hakkari and Tur Abdin regions), and northeastern Syria (Hasakah province).
Over the past century, many Assyrians have emigrated to various parts of the world, including the Caucasus, North America, Europe, and Australia. Events such as the massacres in Diyarbakr, the Assyrian genocide (along with the Greek genocide and Armenian genocide) by the Ottoman Empire during World War I, the Simele massacre in 1933 in Iraq, Arab nationalist policies, the Ba'ath Party in Iraq, the 1979 Iranian Revolution, and the ISIS invasion and occupation of areas in Iraq and Syria have all contributed to the emigration of Assyrians.
It is estimated that the global Assyrian population is around 5 million. Prior to the 1979 revolution, the number of Assyrians in Iran was about 70,000 to 90,000. The majority of Assyrians in Iran are located in cities such as Urmia, Tehran, Shahin Shahr, Salmas, Tabriz, Ahvaz, Hamedan, Kermanshah, Shiraz, Mashhad, Bandar Abbas, Mahshahr, Babolsar, Bandar Anzali, Ferdows and Qazvin.
During the early Islamic centuries, Assyrians made significant contributions to the development of knowledge among Muslims by translating the sciences and knowledge of the ancient world into Arabic and preserving scientific works from antiquity. Some of the greatest scholars and translators of the Islamic Caliphate were from this people. After the Middle Ages, Assyrians came to the attention of the Catholic Church and later the Protestant Church. During the events stemming from World War I and the presence of Russian and Ottoman armies in Azerbaijan, as well as the ethnic and religious conflicts that arose after the assassination of Mar Shimun Benjamin, the leader of the Assyrians, they suffered greatly, and many were forced to migrate elsewhere.
However, in the last century, numerous tragic events have compelled them to migrate to other countries such as the Caucasus, North America, Europe, and Australia. The Diyarbakr massacres, the genocide of Assyrians by the Ottoman Empire, the Simele massacre, nationalist policies, and ISIS attacks are among these factors.
It is estimated that the global Assyrian population is around 5 million, with a significant population in Iran in cities such as Urmia, Tehran, and Tabriz. This nation belongs to the Chaldean Catholic Church, the Assyrian Church of the East, and the Ancient Church of the East, playing an important role in the development of Christianity and sciences throughout different periods. Assyrians, especially during the early Islamic centuries, made significant contributions to the development of knowledge among Muslims through their translations of sciences and knowledge into Arabic.
Ultimately, despite ongoing problems and repeated migrations, Assyrians continue to demonstrate resilience in preserving their culture and identity. This nation remains a descendant of the ancient Assyrian people, descendants of the Prophet Noah, and has maintained its historical and cultural significance even in the face of challenges.</td></tr>
<tr><td>Question</td><td>Answer</td><td>Rate</td></tr>
<tr><td>Which branches of Christianity do the Assyrians predominantly belong to, and what role have they played in the development of Christianity?</td><td>A. Chaldean Catholic Church, Assyrian Church of the East, and Ancient Church of the East; played an important role in the development of Christian and Islamic sciences (correct answer). B. Greek Orthodox Church, Russian Church, and Coptic Church; minor role in translating religious texts (incorrect answer). C. German Protestant Church, American Evangelical Church, and Roman Catholic Church; no significant role (incorrect answer). D. Russian Church, Armenian Church, and Ukrainian Church; significant influence in Christian literature (incorrect answer).</td><td>B
Second part of the answer is incomplete</td></tr>
</table>

Table 14: Translation of sample for rating B for GPT-4o

D Human Evaluation on *PersianMCQ-Instruct*

In our initial attempts to generate Multiple Choice Questions (MCQs) using *PersianMCQ-Instruct*, we found that the questions were quite comprehension-based, required analysis and understanding of the text, and were completely true, grammatically correct, and totally related to the text in 86.8% of cases. However, we also encountered several challenges. When the Persian text contained English words, the generated questions often fell into categories that indicated they were loosely related to the text or contained factual inaccuracies (C and D)(4.7% and 3.2%). We also observed that 5.0% of the outputs fell into category B, meaning that the questions and answers were mostly factual and related to the text, though they may have had some minor grammatical issues or been incomplete."It is noteworthy that a mere 0.3% of the questions were unrelated to the text, a figure so negligible that it can be effectively disregarded. Additionally, This made the evaluation process difficult and time-consuming, as the model tended to produce very complex, lengthy, and very comprehension-based questions and answers.

E Human Evaluation of LLMs on persian MCQs

The evaluation of various LLMs on Persian MCQs before and after fine-tuning reveals significant insights into their performance across different categories (A, B, C, D, and E). Initially, models, including *gemma2-9b-it*, *Llama3.1-8b-Instruct*, and *Mistral-7b-Instruct-v0.3*, struggled with generating questions that met the highest standard of accuracy and relevance. For example, before fine-tuning, 83.5% of outputs for *Mistral-7b-Instruct-v0.3* were predominantly in category E, indicating that the generated questions were completely wrong, misleading, had no clear connection to the source text, or questions/answers were not even generated. Additionally, the presence of non-Persian words in the text often led this model to produce questions categorized as E.The models struggled to accurately identify and label true and false answers, highlighting the challenges in achieving high accuracy and relevance initially.

gemma2-9b-it generated most of its outputs in categories C (32.5%) and B (30.5%), which indicates questions or answers are loosely related to the text but may address topics tangentially, or they mostly factual but potentially incomplete or grammatically flawed responses, or answers that were false. Notably, this model always put the key words in bold without any request from users. Moreover, when non-Persian words were present in the text, *gemma2-9b-it* tended to generate English words in the answers.

Llama3.1-8b-Instruct had 64.5% in category B,this indicates a significant prevalence of grammatical errors, compounded by the model's inability to accurately label true and false answers. and also exhibited specific issues: it often placed the second option as the correct answer, despite being fed that the first option was the true answer. Additionally, when a personality or character name was present in the question, the model frequently failed to follow Persian grammar rules.

Across all three models, we did not have questions in category A, just a few in *gemma2-9b-it*, which means the questions were not factually accurate and directly related to significant concepts from the source text. They were not consistently meaningful and precise, and often had grammatical issues, missing words or questions and answers were not generated.

In fine-tuning, the model resolved the previous issue with non-Persian words in 2 models(*PMCQ-Mistral-7b* and *PMCQ-Llama3.1-8b*) and furthermore, it rectified the challenge of not appropriately identifying and labeling the veracity of answers. Questions were directly related to the text, but they were not comprehension questions like our initial model *PersianMCQ-Instruct*, and most were very similar to the text. The majority of outputs shifted to category A, where questions were accurate, contextually relevant, and free from errors. This shift was especially evident for *PMCQ-Llama3.1-8b* and *PMCQ-Mistral-7b*, where category A became dominant after fine-tuning, with *PMCQ-Llama3.1-8b* achieving 81.5% and *PMCQ-Mistral-7b* achieving 70.5% in category A. This showcases the effectiveness of the fine-tuning process. *PMCQ-Gemma2-9b*, for instance, showed a notable decrease in category B and c, with a post-fine-tuning improvement to 29.5% in category A, indicating that while its pre-fine-tuning outputs were often factual but incomplete, the adjustments resolved these issues, enhancing the clarity and completeness of answers. However this model still had too many problems with non-English words in the text and could not generate accurate questions. It also included too many English words in the questions it generated.

The challenges we faced with *PMCQ-Mistral-7b* were that the generated questions missed all the punctuations, which is important for Persian text. This was not a problem in our initial model, *PersianMCQ-Instruct*. Additionally, compared to our initial model, the questions were less comprehension-based and they are exactly the sentences in the text. This model also had a problem with recognizing numbers.

We minimized categories B, C, D, E and shifted to category A, where questions were accurate, contextually relevant, and free from errors. This reflects the models' improved ability to align generated content with the educational text. The fine-tuned models, such as *PMCQ-Llama3.1-8b*, also displayed a higher frequency of category A outputs, affirming that targeted training refined their understanding and accuracy. This analysis highlights that fine-tuning is essential for transforming LLMs from generating flawed, incomplete, or irrelevant questions into powerful tools capable of producing precise, meaningful, and contextually appropriate MCQs. However, in the model *PMCQ-Gemma2-9b*, we were not as successful; although we shifted to category A, there were still too many outputs in category E.

Overall, the fine-tuning process significantly enhanced the model's performance, transforming it into a more dependable tool for generating meaningful and precise multiple-choice questions (MCQs). While the questions produced are not yet as comprehension-based as those generated by *PersianMCQ-Instruct*, the improvements demonstrate the dataset's effectiveness. This is particularly evident in models like *PMCQ-Mistral-7b* and *PMCQ-Llama3.1-8b*, where the fine-tuning has led to notable advancements in the quality and relevance of the generated MCQs.

E.1 *gemma2-9b-it* (Not Fine-tuned)

- Rating B and C the most frequent with30.5% and 32.5%, indicating that while the questions and answers are mostly factual and related to the text, they may have minor grammatical issues or be incomplete, such as lacking explicit true/false indicators or Questions or answers are loosely related to the text but may address topics tangentially. They may not be correct or may have some serious grammatical issue.

- It also has small number of responses rated as E (28.5%) (demonstrably wrong or misleading content,or not generated),D and A.

E.2 *PMCQ-Gemma2-9b* (Fine-tuned)

- While fine-tuning significantly increases the frequency of Rating A (29.5%), we have made significant strides in enhancing the factual accuracy and precision of our questions and answers. However, this improvement has not yet translated into a reduction in Rating E, which currently stands at 38%. In fact, we have observed an increase in this rating. This suggests that despite the enhancements, the model still produces a notable number of demonstrably wrong or misleading questions, indicating room for further refinement to address these issues.

E.3 *Llama3.1-8b-Instruct* (Not Fine-tuned)

- This model receives a considerable amount of Rating B responses (64.5%). While questions are mostly factual but not fully complete or clear, there is a significant number of D (18.5%) ratings, which suggests that many questions are loosely related to text.

- Before fine-tuning, it had 0% in category A, indicating that its questions were not accurate, contextually relevant, nor free from errors.

E.4 *PMCQ-Llama3.1-8b* (Fine-tuned)

- After fine-tuning, the model's performance shows significant improvement, as there are no outputs rated as B, D, or E. The model predominantly produces outputs rated A (81.5%), indicating high factual accuracy and relevance. There is only a small number of responses rated as C (5.5%) and E (6.5%), suggesting that most questions and answers are highly accurate and well-aligned with the source text, with only a few showing loose relevance or moderate issues.

- This is the best model.

E.5 *Mistral-7b-Instruct-v0.3* (Not Fine-tuned)

- Before fine-tuning, the model's performance is largely poor, with most outputs rated as E (83.5%), indicating a significant number of demonstrably wrong or misleading questions or not generated questions and answers. However, there is a small number of outputs rated as B (2.5%),D(5.0%) and C (9.0%), which means the model can occasionally produce questions that are mostly factual but may have minor issues or loose relevance. This highlights that while the majority of outputs are problematic, some outputs show partial accuracy or moderate quality.

- Before fine-tuning, *Mistral-7b-Instruct-v0.3* has 0% of A, indicating that its questions were not accurate, contextually relevant, nor free from errors.

E.6 *PMCQ-Mistral-7b* (Fine-tuned)

- Fine-tuning results in a significant increase in Rating A (70.5%), which suggests that the model's outputs become more accurate and closely aligned with the source text. The drop in Rating E (9.0%) and other lower categories reflects better performance and reliability.
- For a more detailed analysis, refer to Table 15, which shows the frequency of each rating (A, B, C, D, and E) for each model before and after fine-tuning. This table provides a comprehensive overview of the performance improvements achieved through fine-tuning, highlighting the changes in the distribution of ratings for each model.

	Model Name	**A%**	**B%**	**C%**	**D%**	**E%**
Base	*gemma2-9b-it*	2.5	30.5	32.5	6.0	28.5
	Llama3.1-8b-Instruct	0.0	64.5	9.5	18.5	7.5
	Mistral-7b-Instruct-v0.3	0.0	2.5	9.0	5.0	83.5
Fine-tuned	*PMCQ-Gemma2-9b*	29.5	8.5	11.0	13.0	38.0
	PMCQ-Llama3.1-8b	**81.5**	4.0	5.5	2.5	6.5
	PMCQ-Mistral-7b	70.5	10.0	9.5	1.0	9.0

Table 15: Frequency of Ratings for Each Model Before and After Fine-Tuning