PersianSciQA: A new Dataset for Bridging the Language Gap in Scientific Question Answering

Safoura Aghadavoud Jolfaei
Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
aghadavood@students.irandoc.ac.ir

Azadeh Mohebi
Iranian Research Institute for Information Science and Technology (IranDoc), Tehran, Iran
mohebi@irandoc.ac.ir

Zahra Hemmat
Ferdowsi University of Mashhad, Mashhad, Iran
Hemmat@alumni.um.ac.ir

Abstract

The shortage of specialized datasets hinders the development of Natural Language Processing (NLP) for scientific texts in low-resource languages such as Persian. To address this, we introduce PersianSciQA[1], a large-scale resource of 39,809 question- answer snippet pairs, each containing a question and a scientific answer snippet from a scientific engineering abstract source from IranDoc's 'Ganj'[2] repository, linked by an LLM-assigned relevance score (0-3) that measures how relevant the question is to the content of the accompanying answer snippet. The dataset was generated using a two-stage prompting methodology and refined through a rigorous cleaning pipeline, including text normalization and semantic deduplication. Human validation of 1,000 instances by two NLP researchers confirmed the dataset's quality and a substantial LLM-human agreement (Cohen's kappa coefficient κ=0.6642). To demonstrate its value, we establish baseline benchmarks and show that fine-tuning on PersianSciQA dramatically improves a state-of-the-art model, achieving a Spearman correlation of 0.895 on a blind test set. PersianSciQA provides a crucial new resource to facilitate research in information retrieval and question answering within the Persian scientific domain.

1 Introduction

Natural language processing (NLP) is crucial for navigating and reviewing the vast landscape of new articles published daily by enabling scalable tools for semantic search, question answering, and summarization, to help researchers discover relevant knowledge efficiently. (Hong, 2024; Probierz et al., 2022; Sett & Singh, 2024; Venugopal et al., 2021). As research moves toward knowledge-centered frameworks, effectively utilizing this data becomes imperative. However, while substantial NLP resources exist for high-resource languages, like English (Middha et al., 2024), low-resource languages like Persian remain underrepresented, particularly in specialized scientific domains requiring unique datasets (Jolfaei & Mohebi, 2025; Moniri et al., 2024; Pakray et al., 2025). Despite a growing scientific community, the scarcity of dedicated NLP resources for Persian severely limits the ability to harness its body of scientific literature (Saniee & Arshadi, 2024). To address this gap, we introduce PersianSciQA. Our method employs a hybrid approach, using LLMs for scalable question generation and initial scoring, followed by rigorous human validation. This work aims to harness LLM efficiency to create a large, high-quality resource for the Persian scientific domain, intended to advance research in question answering and information retrieval.

2 Related works

While general Persian resources like PQuAD (Darvishi et al., 2023), FarsiQuAD (ForutanRad et al., 2024), and broader benchmarks such as FaMTEB (Zinvandi et al., 2025) and FarsEval-PKBETS (Shamsfard et al., 2025) exist, they do not fill the specific need for a large-scale scientific corpus.

[1] https://huggingface.co/datasets/safora/persian-scientific-qa

[2] https://en.irandoc.ac.ir/service-product/94

Proceedings of Recent Advances in Natural Language Processing,pages 32–37
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-004

This contrasts sharply with high-resource languages, which benefit from established scientific datasets like SciFact (Wadden et al., 2020), BioASQ (Krithara et al., 2023), SciTail (Khot et al., 2018), and the BEIR benchmark (Thakur et al., 2021), highlighting a critical gap for Persian NLP. Using Large Language Models (LLMs) for dataset creation offers a scalable solution, though their efficacy is debated. Some research indicates LLMs struggle with complex tasks in low-resource settings (Nasution & Onan, 2024; Jadhav et al., 2024), while other studies show LLM-generated data can effectively fine-tune models, sometimes outperforming human-annotated data and reducing costs (Yuen et al., 2025; Li & Cole, 2025). Therefore, while LLMs offer a promising avenue for scalable dataset creation, a validated, large-scale benchmark specifically designed for scientific question answering and relevance ranking in Persian remains a critical and unaddressed need in the current landscape.

3 Dataset Creation Methodology

To construct PersianSciQA, a corpus of 10,846 scientific abstracts from IranDoc's Ganj repository in the engineering field was first used. To generate a diverse and high-quality set of questions, a two-stage process using gpt-4o-mini (OpenAI, 2024), visualized in Figure 1, is designed. Crucially for this work, GPT-4o was also highlighted for its improved capabilities in non-English languages (Yong et al., 2024). Furthermore, the generation of an extensive dataset required a cost-effective solution. In addition to cost, the inference speed of gpt-4o-mini contributed to its selection.

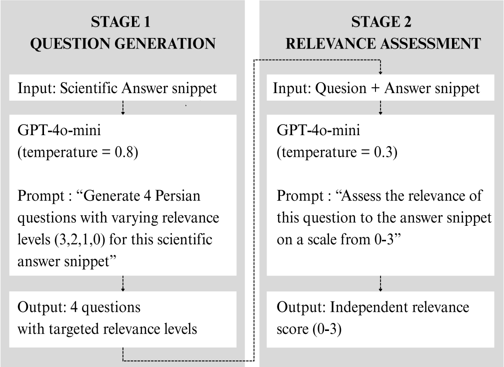
STAGE 1
QUESTION GENERATION
Input: Scientific Answer snippet
GPT-4o-mini
(temperature = 0.8)
Prompt : "Generate 4 Persian questions with varying relevance levels (3,2,1,0) for this scientific answer snippet"
Output: 4 questions
with targeted relevance levels
STAGE 2
RELEVANCE ASSESSMENT
Input: Quesion + Answer snippet
GPT-4o-mini
(temperature = 0.3)
Prompt : "Assess the relevance of this question to the answer snippet on a scale from 0-3"
Output: Independent relevance score (0-3)

Figure 1: Structured Question Generation with Targeted Relevance

At Stage 1 (Generation), the model was prompted with a high temperature (0.8) to generate four distinct Persian questions for each answer snippet; to maximize creativity, we set the temperature to 0.8 in the generation stage. To mitigate self-confirmation bias, at Stage 2 (Assessment) an independent evaluation was done, in which the LLM scored each question-answer snippet pair's relevance using a low, deterministic temperature (0.3), ensuring the generated questions were factually grounded in the source answer snippet.

The resulting 39,883 raw pairs underwent a final refinement pipeline, including Persian text normalization and a two-tier deduplication process (both exact and semantic), to produce the final dataset of 39,809 unique pairs. As detailed in Table 1, the resource exhibits significant scale and lexical diversity, with a question vocabulary of over 17,000 unique words, providing a rich foundation for training robust NLP models.

For instance, it includes direct questions ('What optimization algorithm was used?'; relevance: 3) as well as unrelated ones ('Who funded this research?'; relevance: 0), providing a diverse training signal. The full dataset is publicly available on Hugging Face, upon request.

Metric	**Value**
Total Question- Answer snippet Pairs	39,809
Unique Questions	39,809
Unique Answer snippet	10,235
Avg. Questions per Answer snippet	3.89
Avg. Question Length (words)	14.42
Avg. Answer snippet Length (words)	181.92
Question Vocabulary Size (words)	17,497
Answer snippet Vocabulary Size (words)	86,109

Table 1: Core Statistics of PersianSciQA

A key contribution of this dataset is its balanced four-point relevance score profile, visualized in Figure 2. With a distribution spanning from 'Not Relevant' (19.5%) to 'Highly Relevant' (25.9%), this structure is specifically designed to support the training and evaluation of nuanced relevance ranking models, moving beyond simple binary relevance.

To ensure rigorous and reproducible research, we provide standardized training, validation, and test splits (Table 2). Crucially, these splits are created at the answer snippet level, guaranteeing that all questions related to a single answer snippet are confined to one split. This strategy prevents data leakage and enables fair, robust model evaluation.

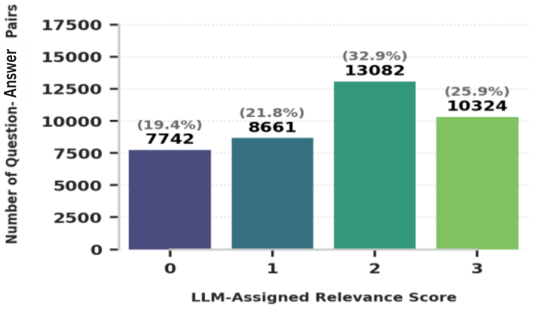

Figure 2: Relevance_Distribution of PersianSciQA Dataset

Metric	Training	Validation	Test
Question- Answer snippet Pairs	31,837	3,968	4,004
Unique Questions	31,837	3,968	4,004
Unique Answer snippet	8,188	1,023	1,024
Percentage of Pairs	80.0%	10.0%	10.0%
Avg. Question/ Answer snippet	3.89	3.88	3.91

Table 2: PersianSciQA Dataset Splits Statistics

4 Experimental Evaluation

To empirically validate PersianSciQA, three key experiments are conducted: (1) evaluating baseline retrieval models, (2) fine-tuning a state-of-the-art model on our dataset to measure its impact on performance and (3) A qualitative review of the fine-tuned model's retrievals.

4.1 Baseline Retrieval Performance

First, the performance of the retrieval model is evaluated using a range of pre-trained embedding models. As shown in Table 3, the experiments revealed a significant finding: top-tier multilingual models like BGE-m3 outperform even the best-specialized Persian-native models on this scientific retrieval task. This provides strong evidence that modern, massively multilingual architectures generalize effectively to specialized low-resource domains, while highlighting the limitations of older or non-retrieval-focused models like ParsBERT.

Model	Type	nDCG@10
BGE-m3	Multilingual	0.4925
multilingual-e5-large	Multilingual	0.4926
Tooka-SBERT	Persian-native	0.3226
paraphrase-MiniLM	Baseline	0.3119
ParsBERT (Base)	Persian-native	0.0507

Table 3: Baseline Retrieval Performance (nDCG@10)

4.2 Fine-Tuning on PersianSciQA

The main contribution is to demonstrate that PersianSciQA is not just a benchmark, but a high-quality resource for improving model performance. the powerful multilingual-e5-large model was selected and fine-tuned on the dataset using a CosineSimilarityLoss function, which is useful for leveraging the graded (0-3) relevance scores. The results, visualized in Figure 3, show a dramatic and consistent improvement in the model's ability to rank relevant documents. The model's Spearman correlation on the validation set was improved from the baseline to 0.892. On the test set, the fine-tuned model achieved a final Spearman correlation of 0.895 and a Pearson correlation of 0.898.

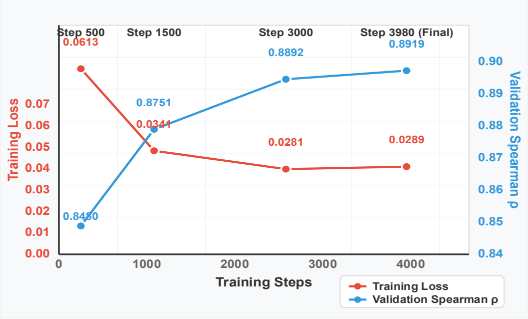

Figure 3: Fine-Tuning Performance on PersianSciQA

This high correlation demonstrates that the model has successfully learned the nuances of semantic relevance within the Persian scientific domain. It confirms that fine-tuning on PersianSciQA provides a significant performance boost, effectively creating a more specialized and accurate retrieval expert. This result is a key contribution, offering the community both a strong new baseline model and definitive proof of the dataset's value for training.

4.3 Qualitative Insights

A qualitative review of the fine-tuned model's retrievals revealed two key behaviors. First, the model demonstrates a strong semantic understanding, consistently retrieving answer snippets that are thematically relevant to the question, even if they do not contain exact keyword matches. Second, the model excels as a "semantic search engine" rather than a direct "answer-finder." It is highly effective at identifying documents about a topic, which can then be passed to a human or a generative model for final answer synthesis. This confirms the model's primary strength lies in high-quality information retrieval, which is the foundational step of any modern Retrieval-Augmented Generation (RAG) system.

5 Human Validation Study

Two NLP researchers independently annotated a randomly selected sample of 1,000 pairs. The initial blind inter-annotator agreement was already substantial. To create the final gold-standard set, the annotators then discussed the few initial disagreements to reach a consensus, a standard practice which resulted in the final adjudicated agreement of Cohen's $\kappa > 0.99$. This two-step process confirms the high quality and clarity of our annotation guidelines and establishes this set as a robust benchmark.

A comparison between the LLM's scores and the gold-standard human labels revealed two key findings. First, there was a substantial LLM-human agreement on relevance, with 75.2% exact score agreement and a Cohen's Kappa of 0.6642. Second, as shown in the confusion matrix (Figure 4), the LLM exhibits a useful conservative bias, consistently underestimating relevance compared to human judges rather than overestimating it. This suggests the model is less likely to assign high relevance incorrectly.

As shown in Figure 5 Linguistic quality assessment by the experts confirmed the high quality of the generated text, with 88.6% of questions deemed clear and grammatically correct. In contrast, only 56.4% of answer snippets were rated as fully acceptable, with readability issues caused by two main factors: formatting challenges inherent in the source documents, which contain mixed-direction Persian and English text, as well as abrupt truncation during the snippet creation process.

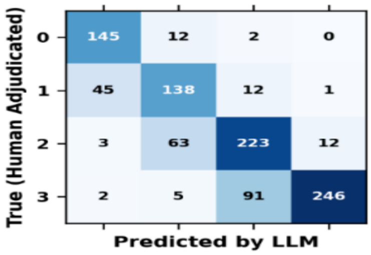

Figure 4: Confusion Matrix of LLM vs. Final Human-Adjudicated Relevance Scores.

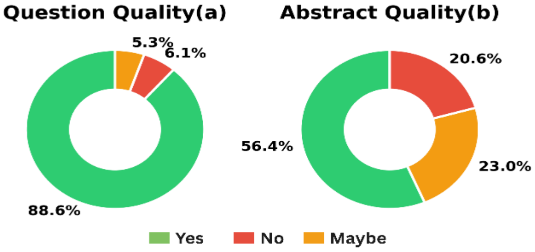

Figure 5: (a,b). question and answer snippet Quality

6 Translation of non-English Terms

The primary application of PersianSciQA is for Question Answering, specifically for training models on answer selection and relevance ranking. As demonstrated in our experiments, it also serves as a robust benchmark for Information Retrieval by treating snippets as mini-documents for semantic search. Furthermore, its rich question diversity facilitates research into paraphrase identification and query understanding within this specialized domain.

7 Conclusion

The PersianSciQA, a new benchmark of 39,809 question-answer snippet pairs was introduced in this research. Its practical value was demonstrated by showing that a model fine-tuned on the data achieves a state-of-the-art 0.895 Spearman correlation. This result confirms that PersianSciQA is not merely a static benchmark but an effective resource for creating expert models for the Persian scientific domain. The dataset's quality was further confirmed through a rigorous human validation process (LLM-human relevance κ=0.6642). While limitations exist, such as the LLM-generated nature, the PersianSciQA is able to provides a valuable resource to spur research and development in Persian scientific information access.

References

Darvishi, K., Shahbodaghkhan, N., Abbasiantaeb, Z., & Momtazi, S. (2023). PQuAD: A Persian question answering dataset. Computer Speech & Language, 80, 101486. https://doi.org/10.1016/j.csl.2023.101486

ForutanRad, J., HourAli, M., & KeyvanRad, M. (2024). Farsi Question and Answer Dataset (FarsiQuAD). Signal and Data Processing, 20(4), 107–120. https://doi.org/10.61186/jsdp.20.4.107

Hong, Z. (2024). Enabling Scientific Information Extraction with Natural Language Processing (Doctoral dissertation, The University of Chicago).

Jadhav, S., Shanbhag, A., Thakurdesai, A., Sinare, R., & Joshi, R. (2024). On Limitations of LLM as Annotator for Low Resource Languages (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2411.17637

Jolfaei, S. A., & Mohebi, A. (2025). A review on persian question answering systems: From traditional to modern approaches. Artificial Intelligence Review, 58(5), 127. https://doi.org/10.1007/s10462-025-11122-z

Khot, T., Sabharwal, A., & Clark, P. (2018). SciTaiL: A Textual Entailment Dataset from Science Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12022

Krithara, A., Nentidis, A., Bougiatiotis, K., & Paliouras, G. (2023). BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Scientific Data, 10(1), 170. https://doi.org/10.1038/s41597-023-02068-4

Li, Z., & Cole, J. M. (2025). Auto-generating question-answering datasets with domain-specific knowledge for language models in scientific tasks. Digital Discovery, 4(4), 998–1005. https://doi.org/10.1039/D4DD00307A

Middha, P., Agarwal, H., Rajput, V., Thakur, A., Singh, S., & Saraswat, S. (2024). Advancing Low Resource Natural Language Processing: Techniques, Applications, and Future Directions. 2024 Second International Conference on Advanced Computing &amp; Communication Technologies (ICACCTech), 337–341. https://doi.org/10.1109/ICACCTech65084.2024.00062

Moniri, S., Schlosser, T., & Kowerko, D. (2024). Investigating the Challenges and Opportunities in Persian Language Information Retrieval through Standardized Data Collections and Deep Learning. Computers, 13(8), 212. https://doi.org/10.3390/computers13080212

Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. IEEE Access, 12, 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809

OpenAI. (2024). GPT-4o-mini - OpenAI API Documentation. OpenAI Platform. https://platform.openai.com/docs/models/gpt-4o-mini

Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. Natural Language Processing, 31(2), 183–197. https://doi.org/10.1017/nlp.2024.33

Probierz, B., Kozak, J., & Hrabia, A. (2022). Clustering of scientific articles using natural language processing. Procedia Computer Science, 207, 3449–3458. https://doi.org/10.1016/j.procs.2022.09.403

Saniee, N., & Arshadi, H. (2024). The Technological Impact of Papers Published by Iranian Institutions: A Scientometric Analysis. International Journal of Information Science and Management (IJISM), Online First. https://doi.org/10.22034/ijism.2024.1990406.1054

Sett, S., & Singh, A. V. (2024). Applying Natural Language Processing in Healthcare Using Data Science. 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 1–6. https://doi.org/10.1109/ICRITO61523.2024.10522196

Shamsfard, M., Saaberi, Z., Hashemi, S. M. H., Vatankhah, Z., Ramezani, M., Pourazin, N., ... & Alipour, S. (2025). FarsEval-PKBETS: A new diverse benchmark for evaluating Persian large language models. arXiv preprint arXiv:2504.14690.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2104.08663

Venugopal, V., Sahoo, S., Zaki, M., Agarwal, M., Gosvami, N. N., & Krishnan, N. M. A. (2021). Looking through glass: Knowledge discovery from materials science literature using natural language processing. Patterns, 2(7), 100290. https://doi.org/10.1016/j.patter.2021.100290

Wadden, D., Lin, S., Lo, K., Wang, L. L., Zuylen, M. van, Cohan, A., & Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims (arXiv:2004.14974). arXiv. https://doi.org/10.48550/arXiv.2004.14974

Yong, Z.-X., Menghini, C., & Bach, S. H. (2024). Low-Resource Languages Jailbreak GPT-4 (arXiv:2310.02446). arXiv. https://doi.org/10.48550/arXiv.2310.02446

Yuen, S., Su, T., Wang, Z., Du, Y., & Sobey, A. J. (2025). Automatic Dataset Generation for Knowledge Intensive Question Answering Tasks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2505.14212

Zinvandi, E., Alikhani, M., Sarmadi, M., Pourbahman, Z., Arvin, S., Kazemi, R., & Amini, A. (2025). FaMTEB: Massive Text Embedding Benchmark in Persian Language (arXiv:2502.11571). arXiv. https://doi.org/10.48550/arXiv.2502.11571