FaMTEB: Massive Text Embedding Benchmark in Persian Language

Erfan Zinvandi[‡†], **Morteza Alikhani**[‡§†], **Mehran Sarmadi**[‡§†], **Zahra Pourbahman**[§]
Sepehr Arvin, Reza Kazemi[‡§], **Arash Amini**[‡§]
[‡]MCINEXT, [§]Sharif University of Technology
*
{e.zeynvandi1376, morteza.alikhani95, mehran.sarmadi99, zahra.pourbahman95,
aamini, reza.kazemi}@sharif.edu, sepehr.arvin@outlook.com

Abstract

In this paper, we introduce a comprehensive benchmark for Persian (Farsi) text embeddings, built upon the Massive Text Embedding Benchmark (MTEB). Our benchmark includes 63 datasets spanning seven different tasks: classification, clustering, pair classification, reranking, retrieval, summary retrieval, and semantic textual similarity. The datasets are a combination of existing, translated, and newly generated (synthetic) data, offering a diverse and robust evaluation framework for Persian language models. All newly translated and synthetic datasets were rigorously evaluated by both humans and automated systems to ensure high quality and reliability. Given the growing adoption of text embedding models in chatbots, evaluation datasets are becoming an essential component of chatbot development and Retrieval-Augmented Generation (RAG) systems. As a contribution, we include chatbot evaluation datasets in the MTEB benchmark for the first time. Additionally, we introduce the novel task of summary retrieval, which is not included in the standard MTEB tasks. Another key contribution of this work is the introduction of a substantial number of new Persian-language NLP datasets for both training and evaluation, many of which have no existing counterparts in Persian. We evaluate the performance of several Persian and multilingual embedding models across a wide range of tasks. This work presents an open-source benchmark with datasets[1], accompanying code, and a public leaderboard[2].

1 Introduction

Text-embedding models aim to generate a semantic vector representation of text, which is helpful in addressing various natural language processing (NLP) tasks such as clustering, classification, semantic textual similarity (STS), information retrieval (IR), and more (Lewis et al., 2020).

To evaluate the performance of models on these tasks, most existing benchmarks are task-specific and fail to assess model capabilities across multiple tasks. For instance, an information retrieval model like Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) may perform well on retrieval tasks but fails to achieve satisfactory results in finding semantic textual similarity (STS). To address this limitation, the massive text embedding benchmark (MTEB) (Muennighoff et al., 2023) introduced a comprehensive evaluation suite across eight diverse NLP tasks, effectively meeting the evaluation needs of text-embedding models for the English language. However, since the primary focus of this benchmark is on English, it does not adequately assess the performance of models in low-resource languages like Persian.

In this work, we introduce **FaMTEB**, a large-scale Persian benchmark for evaluating Persian text-embedding models, enabling users to select the most suitable text-embedding model for their specific tasks. This benchmark comprises 63 datasets across 7 tasks, and we compare the performance of 15 existing Persian or multilingual language models on it. Of these, 24 datasets were pre-existing in Persian, while we contribute 39 new datasets. The newly introduced datasets were developed using three distinct approaches: web-based collection (4 datasets), translation of existing English datasets (16 datasets), and generation through large language models (LLMs) as synthetic datasets (19 datasets). The quality of all datasets has been independently assessed to ensure reliability.

To assess the generalizability of text-embedding models across various NLP tasks, it is essential to conduct a comprehensive evaluation on diverse problems. Since one of the primary and recently highly popular applications of text-embedding

*†These authors contributed equally to this work.

[1]The datasets have been published at https://huggingface.co/MCINext/datasets

[2]Persian space on https://huggingface.co/spaces/mteb/leaderboard

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11441–11468
November 4-9, 2025

models is in retrieval-augmented generation (RAG) systems and chatbots, a portion of the newly curated datasets is dedicated to evaluating these systems. This aspect is explored for the first time within this benchmark. Needless to say that the Persian language is among the low-resource languages in the web and gathering relevant and high quality data is not readily available.

Below we highlight the main contributions of this work:

- Introduction of FaMTEB, a large-scale Persian benchmark for evaluating Persian text-embedding models,
- introduction of a significant number of new Persian datasets in the field of NLP suitable for training and evaluation, some of which have no prior equivalents,
- introduction of the new task of *summary retrieval*, which was not among the 8 tasks included in MTEB,
- introduction of several datasets related to chatbot challenges and RAG systems, which have been included in the MTEB benchmark for the first time.

2 Related work

2.1 Benchmarks

Before the advent of the MTEB (Muennighoff et al., 2023), the evaluation of text embeddings was fragmented across various task-specific and domain-specific benchmarks. Early efforts like the semantic textual similarity benchmarks, including STS-Benchmark (Cer et al., 2017) and SICK (Marelli et al., 2014) datasets, primarily assessed embeddings for their ability to capture semantic relationships between text pairs. While effective in measuring specific aspects, these benchmarks were narrow in scope, focusing on small datasets and failing to represent real-world diversity.

The GLUE (Wang et al., 2018) Benchmark broadened the evaluation landscape by incorporating tasks such as natural language inference, sentiment analysis, and sentence similarity. However, the benchmark is dedicated exclusively to the English language, limiting its applicability to multilingual NLU research and offering limited insight into the generalization capabilities of raw embeddings. Similarly, information retrieval benchmarks like MS MARCO (Nguyen et al., 2016) evaluate embeddings for domain-specific applications, such as search engine optimization. While these benchmarks are invaluable for retrieval tasks, they do not generalize effectively across a wide range of embedding evaluation use-cases.

The fragmented nature of these benchmarks underscore several limitations. First, task diversity is insufficient; most benchmarks aim at specific applications such as similarity, classification, or retrieval without covering clustering or zero-shot classification. Second, inconsistencies in evaluation protocols and metrics make it challenging to compare results across models. Finally, existing benchmarks are not designed for scalability or extensibility; this complicates incorporating new datasets or tasks as embedding techniques evolve.

These limitations led to the creation of MTEB, a unified and comprehensive benchmark that addresses these gaps. MTEB evaluates text embeddings in a wide range of tasks, including semantic similarity, clustering, classification, retrieval, bitext mining, pair classification, and reranking.

Despite the advancements brought by MTEB, its main focus remains on the English language, creating a need for benchmarks tailored to other languages. Although MTEB supports multilingual evaluation, the representation of certain languages and specific linguistic nuances can be limited. To address this shortcoming, new benchmarks inspired by MTEB have been introduced for languages such as Chinese (Xiao et al., 2024), Polish (Poświata et al., 2024), and French (Ciancone et al., 2024). These language-specific benchmarks aim to evaluate the quality of text embeddings in tasks and datasets that reflect the unique linguistic and cultural characteristics of these languages, ensuring broader applicability of text embedding models across the global linguistic landscape.

For the Persian language, besides a limited number of evaluation datasets for isolated tasks, no comprehensive evaluation dataset for text embedding models has been introduced yet. In the STS task, the Farsick (Ghasemi and Keyvanrad, 2021) dataset is available, which is a machine-translated version of the SICK dataset. For the IR task, the multilingual MIRACL (Zhang et al., 2023) dataset supports the Persian language.

2.2 Embedding models

The evolution of text embedding models could be seen as a significant shift from traditional methods like GloVe (Pennington et al., 2014) to more

advanced context-sensitive models. GloVe and Word2Vec (Mikolov et al., 2013) set the foundation for dense word embeddings by representing words as fixed vectors based on co-occurrence statistics. Although these models worked well to capture word-level semantics, they were limited by their static nature, which did not account for polysemy or contextual meaning.

This limitation was addressed with the introduction of ELMo (Peters et al., 2018) and later BERT (Devlin et al., 2019), which offered contextual embeddings by considering the surrounding text. BERT, leveraging the transformer architecture, became a breakthrough by generating deep bidirectional embeddings, capturing richer contextual information. However, BERT's focus on token-level embeddings necessitated fine-tuning for specific tasks, which could be computationally expensive.

To solve this, sentence transformers, like SBERT (Reimers and Gurevych, 2019a) were developed to provide high-quality, task-specific sentence embeddings. These models used the transformer architecture with a Siamese network structure to generate embeddings suitable for tasks such as semantic similarity and information retrieval, making them efficient and versatile for sentence-level understanding.

In recent years, open-source models such as BGE (Xiao et al., 2024), E5 (Wang et al., 2022), and GTE (Li et al., 2023) have been introduced for text embedding, demonstrating strong performance in tasks like STS, retrieval, and clustering in English. In the Persian language, various foundational models such as ParsBERT (Farahani et al., 2020), FaBERT (Masumi et al., 2024), and TookaBERT (SadraeiJavaheri et al., 2024) have been released, which are capable of understanding textual information well; however, they cannot be directly applied to a wide range of tasks. To date, models that perform well in Persian text embedding tasks mainly consist of multilingual networks such as BGE-m3 (Chen et al., 2024), mE5 (Wang et al., 2024), and OpenAI's text embedding models *text-embedding-3* and *text-embedding-4*.

3 The Persian MTEB

3.1 Task definition and evaluation strategy

This benchmark consists of 7 tasks: classification, clustering, semantic textual similarity, retrieval, reranking, pair classification, and summary retrieval. The evaluation methods for all tasks, except summary retrieval, which will be explained later, are detailed in the MTEB (Muennighoff et al., 2023). Figure 1 presents an overview of the tasks and datasets included in FaMTEB.

Summary Retrieval: in this task, we have two input sets. The first set contains a collection of texts, while the second set contains text summaries from the first set. In this evaluation, we search for the text vector from the first set within the second set using the cosine distance. The first retrieved result must be the summary of the corresponding text. F1 score is used as the primary metric for summary retrieval, with accuracy, precision, and recall also being computed to assess the performance.

3.2 New datasets

This work introduces a suite of synthetic, collected, and translated datasets designed to support a broad range of NLP tasks in Persian, spanning STS, dialogue modeling, RAG, sentiment analysis, tone classification, retrieval, QA, and clustering. Synthetic datasets were generated using prompt-based interactions with *GPT-4o-mini*, ensuring coverage of diverse topics, tones, and linguistic phenomena. Collection-based datasets were built through systematic web scraping and labeling pipelines. The translated datasets were produced by applying the Google Translate API to high-quality English benchmarks, enabling consistent evaluation across languages. All newly created datasets underwent a quality evaluation process, discussed in detail in Section 3.3. A detailed description of dataset generation procedures and samples is provided in Appendix A.

3.2.1 BEIR-Fa

A significant portion of our retrieval evaluation data consists of the BEIR benchmark (Thakur et al., 2021) datasets, which have been translated into Persian using the Google Translate service. Similarly, the mMARCO (Bonifacio et al., 2021) evaluation data is also derived by translating the MARCO (Nguyen et al., 2016) dataset using this service. We call this newly generated dataset BEIR-Fa.

3.2.2 Synthetic Persian STS

Since existing Persian datasets are typically translation-based or generated using unsupervised methods—often lacking the quality needed for robust semantic textual similarity (STS) evaluation—we aim to address this gap by construct-

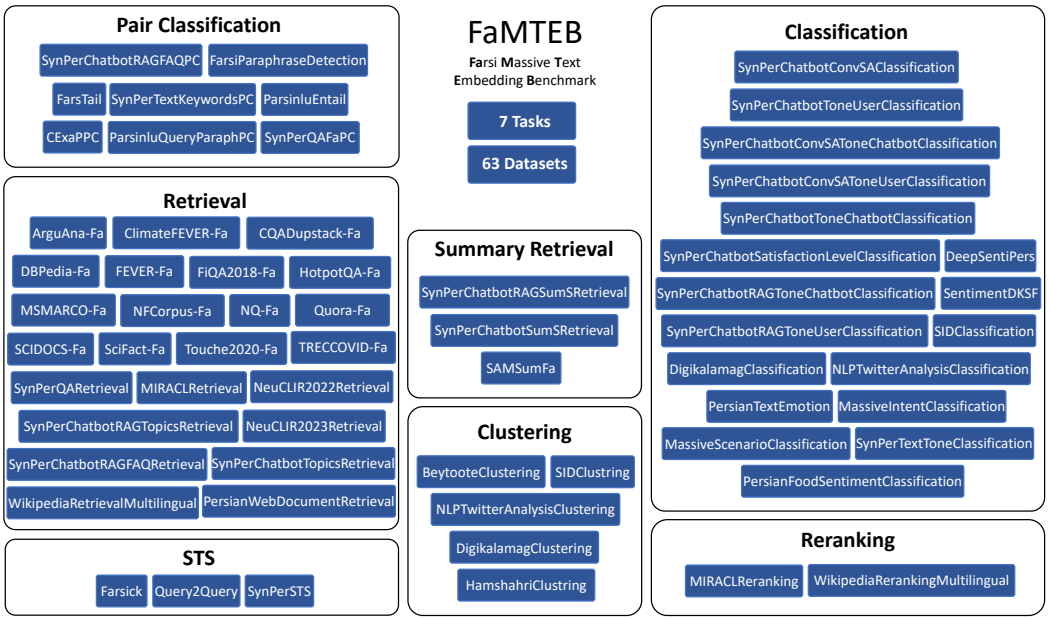


Figure 1: An overview of the FaMTEB evaluation dataset.

ing a high-quality dataset whose annotations are shown to approach human-level performance (see Section 3.3). We present a novel sentence-pair similarity dataset built from diverse Persian texts (Wikipedia, news articles, and informal content) (see Table 5). For each input sentence, we generate five corresponding sentence pairs, each reflecting a different similarity level from 1 to 5, following SemEval-inspired similarity guidelines (Cer et al., 2017). The sentence variants are produced using a prompt-based approach with *GPT-4o-mini*, allowing for fine-grained control over semantic similarity.

3.2.3 Synthetic Persian Chatbot

A multi-purpose chatbot conversation dataset built on 175 topics and 19 tone combinations (from five user/chatbot tones, including formal, casual, childish, aggressive, and street-style). Conversations are generated with a given topic, tone pairing, and one of five user satisfaction levels (including very bad, bad, average, good, and excellent). For each chatbot conversation, a summary of the conversation and related topics have also been generated. From this dataset, we have created several datasets to evaluate different tasks related to chatbots.

3.2.4 Synthetic Persian Chatbot RAG

A multi-purpose chatbot conversation dataset, based on 175 topics and 19 combinations of user and chatbot tones, suitable for RAG-system related tasks. Each generated chatbot conversation in this dataset contains a new user message paired with varying lengths of conversation history, supporting both context-dependent and independent inputs. Each sample also provides a conversation summary, the main topics of the conversation and two FAQ pairs: one that directly answers the user's new message (positive) and one related but not addressing the user's need (negative), making it well-suited for RAG-based applications.

3.2.5 Synthetic Persian Chatbot Conversational Sentiment Analysis

A dataset focusing on user emotions in chatbot interactions. Each sample includes a chatbot conversation, a topic (one of 175 defined topics), tone combination (from 3×3 possibilities: formal, casual, and childish for user and chatbot), an emotion from a set of 9 (i.e., anger, satisfaction, friendship, fear, jealousy, surprise, love, sadness, and happiness), and an intensity level (neutral, moderate, high). Designed for conversational sentiment detection and emotion-aware response modeling

3.2.6 Synthetic Persian Keywords

Similar to the *Synthetic Persian STS dataset*, a collection of texts is selected from various curated Persian websites(see Table 5). In this case, the texts are structured as paragraphs rather than single

sentences.

Using a well-designed prompt with *GPT-4o-mini* (see Figure 4), the desired outputs are generated for each text. These outputs consist of a list of keywords, which represent the key ideas of the text, and a list of non-keywords, which are words in the text that are not considered keywords.

This dataset serves as a comprehensive resource for tasks related to keyword extraction and text analysis in Persian language processing.

3.2.7 Synthetic Persian Tone

We consider the four tones of formal, casual, childish, and literary, and assign each a specific probability. For this dataset, we utilize the paragraphs from the *Synthetic Persian Keywords dataset* (see Figure 4).

For each paragraph, a tone is randomly selected based on the predefined probabilities. Using *GPT-4o-mini*, the paragraph is rewritten in the chosen tone. This process results in the creation of the *Synthetic Persian Tone dataset*, which provides a valuable resource for studying tone transformation and stylistic variations in Persian text.

3.2.8 Synthetic Persian QA

The objective of this dataset is to create a large-scale QA dataset in Persian. To construct this dataset, we select many web pages from curated Persian websites and extract their main content (see Table 5).

For each page, the extracted text is provided to *GPT-4o-mini* along with an appropriately designed prompt (see Figure 4). The model generated multiple question-and-answer pairs based on the content of each page. This process resulted in the creation of the *Synthetic Persian QA dataset*, a valuable resource for question-answering tasks and Persian natural language understanding.

3.2.9 Query to Query

This dataset is constructed based on the assumption that queries yielding similar search results are semantically similar. Queries and results were derived from anonymized query logs and corresponding search results provided by the Zarrebin search engine. Query similarity is quantified as the ratio of overlapping search results to the harmonic mean of the total number of results for each query, computed as:

$$\text{score} = \frac{\text{intersection}}{\frac{(2*Q_1*Q_2)}{Q_1+Q_2}},$$

Similarity scores are then grouped into three levels—unrelated, partially related, and fully related—by applying thresholding on the computed similarity values. Toxic or inappropriate queries are filtered out. The dataset includes many naturally occurring noisy cases, such as misspellings and fragmentary queries, allowing for the evaluation of model robustness in these real-world scenarios.

3.2.10 SID

One approach to dataset creation involves leveraging existing data available on the web. For instance, the MTEB clustering dataset is constructed using category labels from the Arxiv website. Similarly, we adopt this method to develop a clustering dataset for Persian articles. By crawling the **SID**[3] website, which hosts a categorized collection of Persian articles, we curate a dataset comprising of 8 categories. We utilize the title and abstract of each paper, concatenated with two newline characters as input text. Each input text is assigned a single category as the output.

3.2.11 BeytooteClustering

The BeytooteClustering dataset is derived from the collection and categorization of documents from the Beytoote[4] website. Beytoote is a Persian blog that publishes posts on a variety of topics. By crawling the pages of this site, we collect $200,000$ documents and classify them into 22 categories based on the tags present in the URLs. Since each document can be quite lengthy and cover a wide range of topics, we consider only the "summary" tag from the website, which provides a condensed version of the document's content, as the main text. In addition, we remove categories containing fewer than 500 documents to ensure that each category has a sufficient number of entries. The final dataset consists of $95,851$ documents in 19 categories.

3.3 Data Quality Evaluation

3.3.1 Sythetic Data

In this section, we evaluate the synthetic datasets that were generated. To do so, we randomly selected a number of samples from each synthetic dataset for manual tagging by human annotators. Sample sizes ranging from 100 to 600 were selected to correspond with the dataset size. Based on these tags, we calculated a metric for each dataset that reflects the accuracy of the labels generated by

[3] https://sid.ir/

[4] https://www.beytoote.com/

the LLM model. The results are presented in Table 1.

It is important to note that for the STS datasets and the SynPerChatbotSatisfactionLevelClassification dataset, where the labels are ordinal, we used the correlation coefficient (measuring the correlation between the human annotator labels and the labels generated by the LLM) as the evaluation metric. For the other datasets, we used accuracy as the evaluation metric.

Additionally, for retrieval datasets, each sample for tagging consisted of a pair: a query and a related document. The calculated accuracy indicates the percentage of samples where the query and document are indeed related.

For classification, pair classification, and STS datasets, a number of samples were considered for each class. The annotators determined which class each sample truly belongs to, and the metric was calculated based on these annotations.

Another noteworthy point is that each tagging sample was labeled by two annotators. For STS datasets and SynPerChatbotSatisfactionLevelClassification dataset (where the evaluation metric is the correlation coefficient), the final label is the average of the two annotators' labels. For other datasets (where accuracy is the metric), if the two annotators' labels for a sample were identical, the final label for that sample was the same. If the annotators' labels differed, a third annotator determined the final label.

Furthermore, since we evaluated the Query2Query dataset similarly to synthetic STS datasets, its evaluation results are also included in this section and Table 1, even though this dataset was not generated by the LLM.

As shown in Table 1, all datasets exhibit relatively acceptable accuracy, with most datasets achieving an accuracy of over 90

3.3.2 Translated Data

Inspired by the MMARCO paper (Bonifacio et al., 2021) to evaluate translated data, we use the comparison of accuracy measurements in English and Persian using the BM25 metric. In this approach, we index the entire corpus of each dataset using ElasticSearch and retrieve test queries using the BM25 metric. Finally, we report the recall of each dataset for both languages. In Table 4, we observe that the information retrieval accuracy in the translated data does not differ significantly from the accuracy in English, and it can be concluded that the BEIR translated data meets the necessary quality standards.

As another approach to evaluating translated datasets, we adopt the evaluation framework introduced in (Kocmi and Federmann, 2023), which utilizes LLMs to assess translation quality. Using this method, we evaluate translations produced by Google Translate. For comparison, we apply the same procedure to outputs from two LLM models: *GPT-4o-mini* as a closed-source model and *Gemma-3-27B-IT* as an open-source model.

Specifically, for each of the BEIR benchmark datasets translated into Persian, we selected a set of samples and generated translations using *GPT-4o-mini* and *Gemma-3-27B-IT*. Consequently, for each sample, we obtained three distinct translations—one from Google Translate, one from *GPT-4o-mini*, and one from *Gemma-3-27B-IT*. Using the direct assessment method (GEMBA-DA) introduced in the paper (Kocmi and Federmann, 2023), we independently and individually scored each translation on a scale from 0 to 100. We then computed the average scores per BEIR dataset as well as the overall average for each of the three translation approaches. The *GPT-4o* model was also employed as the evaluator in this process.

Table 3 presents the results. As observed, the Google Translate outputs achieve scores above 80 across all datasets, indicating generally high-quality translations. On average, Google Translate received a score of 87.37, which is only slightly lower than the scores of *GPT-4o-mini* and *Gemma-3-27B-IT*, which are 94.65 and 95.08, respectively. These findings demonstrate that translations produced by Google Translate, despite their low cost, maintain good quality that is comparable to those generated by prominent large language models.

3.4 Data Collection

In developing the FaMTEB benchmark, we aimed to ensure diversity across various tasks and topics within each dataset category.

For the classification task, we included a range of problems such as sentiment analysis, tone classification, intent classification, and classification within chat data, using a total of 18 datasets.

In the clustering task, we incorporated 5 datasets covering diverse domains, including news website texts, tweets, and scientific articles.

For the semantic textual similarity (STS) task, the data spans topics such as image captions, user queries, and web content, built from 3 datasets.

Dataset	Evaluation value	Evaluation metric	Dataset type
SynPerChatbotConvSAClassification	93%	Accuracy	Classification
SynPerChatbotToneUserClassification	79%	Accuracy	Classification
SynPerChatbotToneChatbotClassification	89%	Accuracy	Classification
SynPerChatbotRAGToneUserClassification	85%	Accuracy	Classification
SynPerChatbotRAGToneChatbotClassification	86%	Accuracy	Classification
SynPerChatbotConvSAToneUserClassification	94%	Accuracy	Classification
SynPerChatbotConvSAToneChatbotClassification	94%	Accuracy	Classification
SynPerChatbotSatisfactionLevelClassification	0.90	Corr. coef.	Classification
SynPerTextToneClassification	76.0%	Accuracy	Classification
SynPerChatbotSumSRetrieval	100%	Accuracy	Summary Retrieval
SynPerChatbotRAGSumSRetrieval	99%	Accuracy	Summary Retrieval
SynPerChatbotRAGFAQRetrieval	77%	Accuracy	Retrieval
SynPerChatbotTopicsRetrieval	88.5%	Accuracy	Retrieval
SynPerChatbotRAGTopicsRetrieval	93.0%	Accuracy	Retrieval
SynPerQARetrieval	98%	Accuracy	Retrieval
SynPerChatbotRAGFAQPC	85.5%	Accuracy	Pair Classification
SynPerTextKeywordsPC	94.5%	Accuracy	Pair Classification
SynPerQAPC	95.0%	Accuracy	Pair Classification
SynPerSTS	0.911	Corr. coef.	STS
Query2Query	0.695	Corr. coef.	STS

Table 1: Evaluation of synthetic datasets.

	Size (M)	Avg.	Class.	Cluster.	PairClass.	Rerank.	Retriv.	STS	SumRet.
sentence-transformer-parsbert-fa	163	37.93	53.06	**64.83**	70.55	39.9	8.98	55.07	14.71
RoBERTa-WLNI	110	37.97	54.87	58.61	71.1	44.74	9.91	54.92	5.71
BERT-WLNI	110	38.15	54.59	60.32	71.07	45.72	9.82	56.27	6.35
FaBert	124	40.62	60.82	55.18	68.73	50.58	13.92	52.01	9.32
ParsBERT	110	40.63	65.13	56.15	69.14	46.38	9.91	60.97	7.09
paraphrase-multilingual-MiniLM-L12-v2	118	46.62	55.58	58.12	79.9	55.82	23.08	67.24	31.91
LaBSE	471	48.47	62.02	56.56	78.88	55.74	21.2	73.06	47.51
TookaBERT-Base	123	41.17	**65.54**	55.72	70.69	44.18	10.51	61.89	8.29
Tooka-SBERT	353	52.74	61.29	56.45	**87.04**	58.29	27.86	76.41	59.06
GTE-multilingual-base	305	57.14	58.6	57.28	84.57	69.72	41.22	75.75	60.88
multilingual-e5-base	278	57.03	59.97	56.52	84.04	72.07	41.2	74.45	54.58
multilingual-e5-large	560	58.44	61.7	57.19	84.04	74.34	42.98	75.38	56.61
BGE-m3-unsupervised	568	56.6	60.99	59.62	83.07	69.74	38.14	74.47	61.67
BGE-m3	567	59.1	61.74	57.73	85.21	**74.56**	43.38	76.35	61.07
Jina-embeddings-v3	572	**59.28**	62.97	59.15	83.71	61.26	**43.51**	**78.65**	**65.5**

Table 2: Evaluation of various text embedding models on the FaMTEB benchmark. The numbers indicate the percentage of success rate.

BEIR dataset	Google Translate	gpt-4o-mini	Gemma-3-27B-IT
quora	87.03	**96.33**	95.00
dbpedia	80.80	**89.60**	89.23
fiqa	84.88	92.55	**97.28**
arguana	87.45	96.03	**97.08**
scidocs	91.77	**93.23**	92.45
trec-covid	89.44	89.03	**91.13**
nq	80.88	91.35	**98.20**
scifact	90.00	**97.78**	97.60
touche2020	88.30	**97.20**	96.35
msmarco	85.30	**93.93**	91.80
fever	88.15	**96.83**	95.45
cqadupstack	86.90	95.80	**97.88**
hotpotqa	88.70	**96.55**	94.73
climate-fever	89.18	96.95	**97.08**
nfcorpus	91.90	**96.58**	95.03
average	87.37	94.65	**95.08**

Table 3: Scores assigned to translations by three translators—Google Translate, GPT-4o-mini, and Gemma-3-27B-IT across various datasets from the BEIR benchmark.

In the retrieval task, we extended our evaluation beyond the BEIR benchmark—widely used for assessing ranking across diverse topics—by incorporating datasets focused on information retrieval from Wikipedia, web-scale sources, and conversational contexts, resulting in a total of 24 datasets.

For the pair classification task, we evaluate problems such as natural language inference (NLI), paraphrase detection, question answering (QA), and keyword extraction in a pairwise manner, using a total of 8 datasets.

For the reranking task, the evaluation is conducted on Wikipedia pages, and we provide two datasets specifically for this purpose.

In the summary retrieval task, the goal is to retrieve summaries corresponding to dialogues, which include both two-person conversations and interactions between a user and a chatbot. For this task, we provide three distinct datasets.

Language	Evaluation metric	ArguAna	ClimateFEVER	CQADupstack	DBPedia	FEVER	FIQA	HotpotQA	MSMARCO	NFCorpus	NQ	Quora	SCIDOCS	SciFact	Touche2020	TRECCOVID
Persian	ndcg@10	42.93	15.32	26.58	25.45	50.84	17.44	46.53	38.27	33.25	22.02	61.71	14.06	65.31	26.49	56.9
English	ndcg@10	47.16	18.61	32.52	32.01	64.93	25.36	60.22	47.68	34.28	32.60	88.77	16.46	69.06	34.70	68.80
Persian	recall@100	93.67	34.29	54.63	35.83	82.98	46.94	65.11	33.02	26.50	60.78	86.03	32.31	89.7	47.18	10.50
English	recall@100	95.16	40.93	62.13	43.47	92.15	54.88	76.30	45.02	26.02	78.28	97.69	36.75	91.92	56.09	11.73

Table 4: Accuracy evaluations conducted on English and Persian languages using the BM25 metric.

	Synthetic Persian STS	Synthetic Persian Keywords / Tone	type
wikipedia	5000	1000	wiki
wikishia	500	500	wiki
wiki.ahlolbait	500	500	wiki
hawzah	500	500	wiki
yjc	1000	1000	news
tasnim	500	1000	news
tabnak	500	1000	news
beytoote	500	1000	news
varzesh3	500	1000	news
isna	1000	1000	news
mehrnews	1000	1000	news
asriran	1000	1000	news
khabaronline	500	500	news
ecoiran	500	500	news
hamshahrionline	0	500	news
donyaeeqtesad	0	500	news
alibaba	500	500	article
khanesarmaye	500	0	article
digiato	500	500	article
ninisite / article	500	500	article
zoomit	500	500	article
bigbangpage	500	500	article
namnak	500	500	article
hamgardi	0	500	article
ninisite / discussion	500	2000	informal
voolak	500	1500	informal
doctoreto	500	500	others
taaghche	500	500	others
virgool	500	0	others
soft98	0	500	others
vipofilm	0	500	others

Table 5: The number of documents selected for each dataset from various websites.

4 Results

4.1 Models

The models we evaluate are categorized into three main groups. The first group consists of multilingual text embedding models that support Persian, including multilingual-e5 (Wang et al., 2024), BGE-m3 (Chen et al., 2024), GTE-multilingual (Zhang et al., 2024), paraphrase multilingual, paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019b), LaBSE (Feng et al., 2022), and Jina (Sturua et al., 2024). The second group consists of text embedding models specifically trained for Persian, such as BERT-WLNI, RoBERTa-WLNI, sentence-transformer-parsbert, and Tooka-SBERT. The third group includes base model transformers trained in Persian, namely ParsBERT (Farahani et al., 2020), FaBERT (Masumi et al., 2024), and TookaBERT (SadraeiJavaheri et al., 2024).

4.2 Analysis

In this section, we evaluate the models introduced in Section 4.1 on the FaMTEB benchmark. The results of our experiments are presented in Table 2. It can be observed that the Jina model achieves the highest accuracy on the FaMTEB benchmark in terms of average accuracy. This model demonstrates the best performance across three tasks: STS, retrieval, and summary retrieval, compared to other models. Additionally, the BGE-m3 model achieves the highest accuracy in the rerank task and exhibits an accuracy close to that of the Jina model in the retrieval task. Therefore, we identify this model as a tailored model for the information retrieval task. Moreover, in the classification, clustering, and pair classification tasks, we observe that the sentence-transformer-parsbert, TookaBERT, and Tooka-SBERT models, which are trained specifically for the Persian language, achieve the highest accuracies, respectively.

5 Conclusion

In this study, we presented FaMTEB, a comprehensive benchmark for evaluating Persian text embeddings, building upon the massive text embedding benchmark (MTEB). Our contribution spans several dimensions, including the introduction of 63 datasets across seven diverse tasks, the addition of a novel task (summary retrieval), and the creation of numerous new Persian-language NLP datasets for training and evaluation. By integrating datasets related to chatbot challenges and RAG systems into MTEB for the first time, we provide a tailored evaluation framework that aligns with emerging applications of text embedding models. The performance evaluation of Persian and multilingual embedding models further underscores the benchmark's utility. This open-source benchmark, encompassing datasets, code and a public leaderboard, establishes a robust foundation for advancing NLP research in the Persian language. We hope FaMTEB inspires further innovation and enhances the development of more effective Persian-language models.

As part of future work, we plan to reduce the size of selected BEIR datasets and explore the

use of closed-source LLMs such as *GPT-4o-mini*, as well as powerful open-source alternatives like *Gemma-3-27B-IT*, for translation purposes. This will improve the quality of dataset translations. Additionally, we intend to expand the benchmark by including results from a wider range of embedding models, particularly those based on LLMs, to enable a more comprehensive evaluation across architectures and training paradigms.

Limitations

Human-annotated datasets Despite advancements in Persian natural language processing (NLP), the availability of human-annotated datasets remains limited. The scarcity of such data poses a significant challenge in training and evaluating high-quality models across various NLP tasks. Addressing this limitation requires extensive efforts to construct diverse and representative datasets tailored to different linguistic problems. This challenge is particularly evident in tasks such as semantic textual similarity, reranking, and summarization, where high-quality labeled data is essential for achieving reliable performance.

Closed-source models Due to the high costs associated with evaluating text embedding models that provide their APIs on a paid basis, such as *text-embedding-3-large*, as well as the resource-intensive nature of serving and evaluating some other open-source text embedding models, we have not yet included certain models in the leaderboard. We are gradually adding these models over time.

Machine translation Due to the large size and large number of samples in the BEIR datasets, we relied on Google Translate for translation, as it offers a highly cost-effective solution. However, this choice may have limited the overall translation quality.

Annotators A potential limitation of this work is that the annotators for the dataset were the paper authors themselves. To mitigate any potential bias, all annotation guidelines and instructions were collaboratively developed and shared among the annotators to ensure consistency and a unified understanding of the task.

Data leakage The concern regarding data leakage is indeed important when constructing evaluation datasets. Many LLMs today have been trained on vast portions of the web, and therefore, when such datasets are used, care must be taken to ensure that models have not already been exposed to them during training. In the case of our synthetic datasets, we rely on existing web data as a form of prior knowledge and then employ LLMs to transform this material into task-specific datasets. While the likelihood of an LLM having encountered the exact newly generated data is nearly zero, it remains possible that the model has previously been exposed to similar sentence structures or distributional patterns, given its role in generating the dataset.

Ethics

License In conducting this research, we carefully considered the ethical and legal aspects of using LLMs. Specifically, the outputs generated from OpenAI models and Google's Gemma models were utilized in a manner consistent with their respective usage guidelines and licenses.

The use of OpenAI models is governed by the OpenAI Terms of Use,[5] which restricts harmful or unlawful applications and specifies conditions for redistribution of generated outputs. Similarly, the Gemma family of models is released under an open license intended to promote responsible research and development,[6] while also requiring adherence to standards that prevent misuse.

In this work, outputs from these models were applied strictly for academic and non-commercial purposes. The generated content was not used in ways that could cause harm, misrepresentation, or violation of intellectual property rights. All uses were in accordance with the attribution requirements and ethical norms of transparency in AI-assisted research.

References

Mehdi Allahyar. 2020. farsi news.

Hossein Amirkhani, Mohammad AzariJafari, Zohreh Pourjafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, and Azadeh Amirak. 2020. Farstail: A persian natural language inference dataset. *CoRR*, abs/2009.08820.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto A. Lotufo, and Rodrigo Frassetto Nogueira. 2021. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality,

[5] https://openai.com/policies/row-terms-of-use/

[6] https://ai.google.dev/gemma/terms

multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv:2405.20468*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53:3831 – 3847.

Hamed Feizabadi. 2023. nlp twitter analysis.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Wikimedia Foundation. 2023. Wikimedia downloads.

Ali Ghasemi. 2022. farsi paraphrase detection.

Zahra Ghasemi and Mohammad Ali Keyvanrad. 2021. Farsick: A persian semantic textual similarity and natural language inference dataset. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 194–199.

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for persian. *ArXiv*, abs/1801.07737.

Seyed Ali Mir Mohammad Hosseini. 2022. Persian text emotion.

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

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *Preprint*, arXiv:2404.08071.

Dawn J Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the trec 2022 neuclir track. *ArXiv*, abs/2304.12367.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mostafa Masumi, Seyed Soroush Majd, Mehrnoush Shamsfard, and Hamid Beigy. 2024. Fabert: Pre-training bert on persian blogs. *ArXiv*, abs/2402.06617.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

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

Reyhaneh Sadeghi, Hamed Karbasi, and Ahmad Akbari. 2022. Exappc: a large-scale persian paraphrase detection corpus. In *2022 8th International Conference on Web Research (ICWR)*, pages 168–175.

MohammadAli SadraeiJavaheri, Ali Moghaddaszadeh, Milad Molazadeh, Fariba Naeiji, Farnaz Aghababaloo, Hamideh Rafiee, Zahra Amirmahani, Tohid Abedini, Fatemeh Zahra Sheikhi, and Amirmohammad Salehoof. 2024. Tookabert: A step forward for persian nlu. *Preprint*, arXiv:2407.16382.

Javad Pourmostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. *ArXiv*, abs/2004.05328.

Aryan Shekarlaban and Pooya Mohammadi Kazaj. 2023. Hezar: The all-in-one ai library for persian. `https://github.com/hezarai/hezar`.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

1	The two sentences are completely dissimilar but the writing structure of the two texts can be similar.
2	The two sentences are not equivalent but share the same topic.
3	The two sentences are roughly equivalent, but some important information differs/missing.
4	The two sentences are mostly equivalent, but some unimportant details differ
5	The two sentences are completely equivalent, as they mean the same thing.

Table 6: Similarity score definition based on (Cer et al., 2017).

Erfan Zinvandi, Morteza Alikhani, Zahra Pourbahman, Reza Kazemi, and Arash Amini. 2024. Persian web document retrieval corpus. In *2024 12th Iran Workshop on Communication and Information Theory (IWCIT)*, pages 1–3.

A Datasets Details

A.1 Synthetic Persian STS

To the best of our knowledge, no independently constructed dataset exists for the STS task in Persian. Available datasets are typically derived from translations of English datasets or created using unsupervised methods. For instance, the FarSICK dataset is a translation of the SICK dataset. To address this lack of datasets in Persian, we propose constructing a new dataset using an LLM.

The first step involves compiling a collection of sentences with a suitable distribution of diverse topics. To achieve this, we sample data from various websites covering topics such as Wiki, news, articles, informal, and others. For generating each sample, the documents from each site are segmented at the sentence level using the *nltk* library. Then, sentences that meet the following criteria are selected: they must contain between 35 and 200 characters in length, and at least 30 percent of their characters must be in Persian. These conditions ensure that each sentence is complete, meaningful, and not overly ambiguous.

Subsequently, we employ a prompt-based method, inspired by the rules for determining sentence similarity levels proposed in SemEval (Cer et al., 2017). However, instead of the six levels of similarity used in SemEval, we define five levels, with the final score being a discrete value ranging from 1 to 5 (see Tabel 6). Finally, we provide samples with the given prompt to *GPT-4o-mini*[7] to generate STS data.

[7]https://chat.openai.com/

A.2 Synthetic Persian chatbot

Given the increasing prevalence and usage of diverse chatbots, along with the various challenges in chatbot systems that text embeddings and language models can address, the creation of datasets specific to chatbot-related challenges has become highly significant.

The goal of constructing the synthetic Persian chatbot dataset is to generate multiple datasets focused on chatbot-related challenges. The dataset generation process involves several steps. First, 175 distinct topics are identified as conversation subjects for users interacting with chatbots. These topics are selected to be diverse, covering various chatbot types and scenarios, with the final number of samples for each topic determined accordingly.

Five distinct tones are considered for the conversations including formal, casual, childish, aggressive, and street-style. Additionally, 19 combinations of user and chatbot tones are designed by pairing these tones. Probabilities are assigned to each combination, ensuring that the total probabilities across the combinations equaled one.

Five levels of user satisfaction are defined: very bad, bad, average, good, and excellent. For each topic, samples are generated to correspond to each satisfaction level, with the number of samples for the "excellent" level being twice as much as other levels.

For each sample, a tone combination for the user and chatbot is selected based on the assigned probabilities. Each sample included a topic, a user satisfaction level, a user tone, and a chatbot tone. Using these specifications, a prompt is created and provided to *GPT-4o-mini* to generate a conversation between the user and the chatbot. The generated conversation adheres to the specified topic, user tone, chatbot tone, and user satisfaction level. Additionally, the prompt instructs the model to produce supplementary outputs alongside the conversation, such as a summary of the dialogue and the key topics discussed.

Finally, from this comprehensive dataset, several specialized datasets are derived, each tailored for specific purposes.

A.3 Synthetic Persian chatbot RAG

Unlike the Synthetic Persian chatbot dataset, where each sample contained a complete conversation between the user and the chatbot, in the synthetic Persian chatbot RAG dataset, the chat is not neces-

sarily complete. During the conversation between the user and the chatbot, one of the user's messages is considered as the new user input, and the output should respond to this message. This message may either be a follow-up, i.e., it dependents on previous interactions, or it may not be contextually linked to prior conversations, making the dataset particularly suitable for RAG systems.

Each sample consists of the user's new message along the history of previous user-chatbot conversations. This design aligns with the requirements of RAG systems. The dataset shares components with the synthetic Persian chatbot dataset, such as using the same 175 topics as conversation subjects and applying the 19 combinations of user and chatbot tones from the previous dataset. However, this dataset does not include the final user satisfaction level.

To construct the dataset, a number from the set $\{0, 2, 4, \ldots, 20\}$ is randomly chosen to determine the number of historical messages in the conversation. Each number in the set has an assigned selection probability. Based on the specified parameters, a suitable prompt is crafted and provided to *GPT-4o-mini* to generate the output. This output includes the desired conversation (conversation history plus the new user message), a summary of the conversation, the main topics of the conversation, and two FAQ-related samples. The positive FAQ sample is a question-answer pair that resolved the user's need expressed in the new message, while the negative FAQ sample is a question-answer pair related to the conversation but does not address the user's need in the new message.

The resulting dataset designed with this structure, is further used to create several specialized datasets, each tailored for different applications.

A.4 Synthetic Persian chatbot conversational sentiment analysis

The primary objective of this dataset is to create a sentiment analysis dataset focusing on user emotions during interactions with chatbots. The dataset incorporates nine distinct emotions: anger, satisfaction, friendship, fear, jealousy, surprise, love, sadness, and happiness. Each emotion is assigned a selection probability, with the total probabilities summing to 1.

Three tones are considered for both the user and the chatbot: formal, casual, and childish. This results in nine possible tone combinations, each with a specific selection probability. Each sample in the dataset includes a topic chosen from the 175 predefined topics, one of the nine tone combinations for the user and chatbot, one of the nine predefined emotions, and an intensity level for the selected emotion, categorized into three levels: neutral, moderate, and high.

Using these specifications, an appropriate prompt is crafted and provided to *GPT-4o-mini*. The output consistes of a conversation between the user and the chatbot that adheres to the required topic, tone, emotion, and emotion intensity level.

The generated dataset is further utilized to create additional specialized datasets tailored for specific sentiment analysis applications.

A.5 Query to query

Logically, the queries yielding similar results in a search engine are inherently similar. Thus, the degree of overlap in search results between two queries roughly measures their similarity. The data for this dataset is collected using anonymous user queries and search results from the *Zarrebin* search engine[8], a platform designed for the Persian language. The similarity score between query pairs is normalized to account for differences in the number of responses. The similarity score is calculated by dividing the intersection of responses by the harmonic mean of the total number of responses for the two queries:

$$\text{score} = \frac{\text{intersection}}{\frac{(2 * Q_1 * Q_2)}{Q_1 + Q_2}},$$

where "intersection" represents the number of common responses, while Q_1 and Q_2 refer to the total number of responses for the first and second queries, respectively.

Queries containing harmful or toxic content are filtered for integrity of the dataset. The resulting scores are categorized into three levels: unrelated, partially related, and fully related. Thresholds of 0.2 and 0.4 are determined empirically. The dataset is beneficial as it includes query pairs with spelling errors or incomplete sentences, requiring the model to handle deficiencies. Furthermore, the dataset includes queries requiring meaning comprehension, such as when users search for a song in different ways, expecting similar results. This highlights the model's ability to assess semantic relationships despite phrasing differences.

[8] https://zarebin.ir/

B Datasets Categories

In this section, we review the complete set of datasets used in the FaMTEB benchmark. Table 7 provides the number of data for each dataset.

Figure 6 presents the similarity chart of different datasets. To construct this chart, we first randomly selected 100 samples from each dataset and computed the average vector of these samples. Next, we calculated the cosine similarity between the resulting vectors for the datasets, multiplied the values by 100, and rounded them to obtain the final chart.

For vector representation, we utilized the Jina-embeddings-v3 model, which demonstrated the highest performance among the evaluated models in both the STS task and overall average performance (see Table 2).

As observed, the FaMTEB datasets exhibit a wide range of similarity values. In general, most similarity scores between datasets are relatively low, indicating the diverse nature of the FaMTEB datasets.

We have also provided a sample from each of the different datasets available in this benchmark in Figures 7, 8, 9, 10, 11, 12, and 13.

The distribution of sources used to construct the *synthetic Persian STS*, *synthetic Persian keywords*, and *synthetic Persian tone* datasets is detailed in Table 5. Furthermore, the set of prompts employed for generating the synthetic datasets is depicted in Figures 4, 5, and 2.

B.1 Classification

DigikalamagClassification (Farahani et al., 2020) The Digikala Magazine dataset (DigikalamagClassification) consists of 8,515 articles scraped from the Digikala Online Magazine. The dataset is organized into seven distinct classes: Video Games, Shopping Guide, Health & Beauty, Science & Technology, General, Art & Cinema, and Books & Literature.

NLPTwitterAnalysisClassification (Feizabadi, 2023) The NLPTwitterAnalysisClassification is a classification dataset that categorizes various tweets into 26 distinct classes based on the topics they discuss.

SentimentDKSF (Shekarlaban and Kazaj, 2023) This dataset is composed of comments collected from the Digikala [9] and SnappFood [10] websites. Each comment is annotated with one of three labels: Positive, Negative, or Neutral, indicating the user's level of satisfaction.

PersianTextEmotion (Hosseini, 2022) PersianTextEmotion is a Persian sentiment analysis dataset containing 6,948 sentences, each annotated with one of six distinct emotions: joy, sadness, anger, disgust, fear, or surprise.

PersianFoodSentimentClassification (Farahani et al., 2020) The PersianFoodSentimentClassification dataset consists of 70,000 user comments collected from SnappFood, an online food delivery platform. The dataset is designed for polarity classification and includes two sentiment labels: (0) Happy and (1) Sad. This dataset provides a comprehensive resource for sentiment analysis, particularly in assessing customer satisfaction and experience within the context of online food delivery services.

DeepSentiPers (Sharami et al., 2020) The DeepSentiPers dataset is an enhanced and balanced version of the SentiPers dataset (Hosseini et al., 2018), containing 12,138 user opinions on digital products. It is annotated with five sentiment classes: two positive (Happy and Delighted), two negative (Furious and Angry), and one Neutral class. This dataset can be utilized for both multi-class and binary sentiment classification tasks. For binary classification, the Neutral class is excluded, focusing the analysis on positive and negative sentiments.

MassiveIntentClassification (FitzGerald et al., 2023) This dataset comprises a collection of Amazon Alexa virtual assistant utterances, each annotated with its corresponding intent. One of 60 possible intents is assigned as the label for each user utterance. The dataset is multilingual, supporting 51 languages, including Persian.

MassiveScenarioClassification (FitzGerald et al., 2023) This dataset comprises a collection of Amazon Alexa virtual assistant utterances, each annotated with its corresponding scenario. One of 60 possible scenarios is assigned as the label for each user utterance. The dataset is multilingual, supporting 51 languages, including Persian.

SynPerChatbotConvSAClassification (Some Emotion) For each of the nine emotions in the *Synthetic Persian Chatbot Conversational Sentiment Analysis*, a separate classification dataset was created. The samples for each dataset were selected from the *Synthetic Persian Chatbot Conversational Sentiment Analysis Dataset*, focusing on conversations that exhibit the specific emotion of interest.

[9] digikala.com
[10] snappfood.ir

Each sample represents a conversation between the user and the chatbot, and is labeled with one of two classes:

- **Negative**: When the intensity level of the emotion is zero.
- **Positive**: When the intensity level of the emotion is moderate or high.

SynPerChatbotToneUserClassification This is a classification dataset derived from the *Synthetic Persian Chatbot*, which classifies the user's tone in the conversation between the user and the chatbot.

SynPerChatbotToneChatbotClassification This is a classification dataset derived from the *Synthetic Persian Chatbot*, which classifies the chatbot's tone in the conversation between the user and the chatbot.

SynPerChatbotRAGToneUserClassification This is a classification dataset derived from the *Synthetic Persian Chatbot RAG*, which classifies the user's tone in the conversation between the user and the chatbot.

SynPerChatbotRAGToneChatbotClassification This is a classification dataset derived from the *Synthetic Persian Chatbot RAG*, which classifies the chatbot's tone in the conversation between the user and the chatbot.

SynPerChatbotConvSAToneUserClassification This is a classification dataset derived from the *Synthetic Persian Chatbot Conversational Sentiment Analysis*, which classifies the user's tone in the conversation between the user and the chatbot.

SynPerChatbotConvSAToneChatbotClassification This is a classification dataset derived from the *Synthetic Persian Chatbot Conversational Sentiment Analysis*, which classifies the chatbot's tone in the conversation between the user and the chatbot.

SynPerChatbotSatisfactionLevelClassification This is a classification dataset derived from the *Synthetic Persian Chatbot Dataset*, which classifies the user's satisfaction level in the conversation between the user and the chatbot.

SynPerTextToneClassification This is a classification dataset derived from the *Synthetic Persian Tone Dataset*. The samples consist of rewritten texts, and the classes correspond to the tones of the rewritten texts.

SIDClassification As detailed in Section 3.2.10, the SIDClassification dataset is derived from the SID[11] website. In this dataset, the titles and abstracts of articles are used as input texts, which are classified according to the website's predefined category labels, with these categories serving as the class labels.

B.2 Clustering

BeytooteClustering As explained in Section 3.2.11, BeytooteClustering is a clustering dataset collected from the Beytoote[12] website. The BeytooteClustering dataset consists of 19 categories, as detailed in Table 8.

DigikalamagClustering This dataset is the same as the DigikalamagClassification dataset described in Section B.1, with the distinction that it is used here for the clustering task.

HamshahriClustring (Allahyar, 2020) The Hamshahri dataset is a subset of the Farsi-news dataset, extracted from the RSS feeds of two prominent Farsi news agency websites: Hamshahri and RadioFarda. Each entry in this dataset consists of the title, summary, URL, and tags of the respective news page. For the Hamshahri dataset specifically, the title and summary are concatenated to form the input text, while the tag associated with each page is used as the classification label.

NLPTwitterAnalysisClustering This dataset is the same as the NLPTwitterAnalysisClassification dataset described in Section B.1, with the distinction that it is used here for the clustering task.

SIDClustring As described in Section 3.2.10, SIDClustring is a clustering dataset sourced from the SID[13] website. In this dataset, the title and abstract of articles serve as input texts, which are categorized based on the website's predefined category labels.

B.3 STS

Farsick (Ghasemi and Keyvanrad, 2021) FarSick is a Semantic Textual Similarity (STS) dataset designed for the Persian language. The dataset comprises approximately 10,000 sentence pairs, each meticulously annotated for semantic relatedness, meaning alignment, and entailment relations between the sentence pairs. FarSick was developed by translating and adapting the sentence pairs from the original SICK dataset, ensuring relevance and applicability to the Persian linguistic context.

SynPerSTS As described in 3.2.2, we utilize GPT-4o-mini to generate synthetic data for the Semantic Textual Similarity (STS) task.

[11] https://www.sid.com/

[12] https://www.beytoote.com/

[13] https://www.sid.com/

Query2Query As detailed in Section 3.2.9, this is a Semantic Textual Similarity (STS) dataset comprising search engine queries, categorized into three levels of similarity. The distribution of labels is shown in Figure 3.

B.4 Summary Retrieval

SynPerChatbotSumSRetrieval This dataset is derived from the *Synthetic Persian Chatbot Dataset*. It is designed to evaluate the model's ability to retrieve summaries of conversations between the user and the chatbot.

SynPerChatbotRAGSumSRetrieval This dataset is derived from the *Synthetic Persian Chatbot RAG Dataset*. It is designed to evaluate the model's ability to retrieve summaries of user-chatbot conversations, including conversations that may not have been completed.

SAMSumFa The SAMSum dataset is a dataset designed for abstractive dialogue summarization. It is composed of real-world-like chat conversations along with human-written summaries. SAMSumFa is a translation of the SAMSum dataset using the exact method proposed for BEIR in Section 3.2.1.

B.5 Retrieval

ArguAna-Fa ArguAna-Fa retrieves the best counterargument to an argument. This dataset aids in assessing the relevance and quality of counterarguments provided in debates or discussions.

ClimateFEVER-Fa ClimateFEVER-Fa verifies climate claims using the FEVER Wiki corpus, with claims as queries and evidence retrieval. The dataset is used to evaluate the effectiveness of systems in fact-checking climate-related statements based on available evidence.

CQADupstack-Fa CQADupstack-Fa is a community question-answering dataset with queries from 12 StackExchange subforums, evaluated using retrieval of duplicate queries, reporting mean scores in BEIR. It focuses on identifying and retrieving relevant information based on user questions across a variety of topics.

DBPedia-Fa DBPedia-Fa is an entity retrieval dataset with heterogeneous queries, focusing on retrieving entities from the English DBpedia corpus. It is valuable for tasks that involve querying knowledge bases and retrieving structured data based on user inputs.

FEVER-Fa FEVER-Fa dataset supports automatic fact-checking by retrieving evidence from pre-processed Wikipedia abstracts using original paper splits as queries. This dataset is particularly useful for evaluating models that perform fact-checking tasks, ensuring they can retrieve reliable sources of evidence.

FiQA2018-Fa FiQA2018-Fa focuses on opinion-based QA, using financial data from StackExchange Investment posts. The dataset evaluates the ability of systems to answer subjective questions related to financial markets, opinions, and trends.

HotpotQA-Fa HotpotQA-Fa requires obtaining the correct answer to a question by reasoning in multiple paragraphs. It challenges models to combine information from several documents to arrive at accurate answers, thus testing systems' multi-hop reasoning abilities.

MSMARCO-Fa MSMARCO-Fa is an information retrieval dataset comprising queries from Bing logs and various web documents, where each query is associated with both relevant and irrelevant documents. This dataset is used to evaluate information retrieval systems, particularly their ability to rank documents by relevance to the query.

NFCorpus-Fa NFCorpus-Fa includes queries from NutritionFacts and an annotated medical corpus from PubMed. The dataset supports research in information retrieval in the medical domain, where accurate and relevant information is critical for health-related queries.

NQ-Fa NQ-Fa consists of a set of Google search engine queries and their corresponding answers from Wikipedia, which have been human-annotated. This dataset is designed for training and evaluating models on natural question answering tasks.

Quora-Fa Quora-Fa consists of the Quora corpus, where duplicate queries have been identified. It is primarily used to assess the ability of systems to detect duplicate questions and provide consistent answers.

SCIDOCS-Fa SCIDOCS-Fa is a citation prediction task aimed at identifying directly cited scientific articles based solely on the title of the citing article. The dataset facilitates research into automatic citation prediction and understanding of scientific knowledge connections.

SciFact-Fa SciFact-Fa dataset is a scientific fact verification benchmark that evaluates the ability of models to verify claims using evidence from the scientific literature, focusing on claim-evidence alignment and factual correctness. It is critical for fact-checking tasks in the scientific domain.

Touche2020-Fa Touche2020-Fa is a dataset designed for argument retrieval tasks. The goal is to identify and retrieve relevant arguments to counter or support claims based on a given premise. It evaluates models' ability to engage in argumentative reasoning by considering both the quality and relevance of retrieved arguments.

TRECCOVID-Fa TRECCOVID-Fa is an ad-hoc search challenge using the July 16, 2020 CORD-19 dataset with cumulative judgments and query descriptions from the original task. This dataset is used for evaluating information retrieval systems on COVID-19-related content.

Note: All of these datasets have been translated from the original BEIR datasets to the Persian language as explained in Section 3.2.1.

SynPerChatbotRAGFAQRetrieval This dataset is derived from the *Synthetic Persian Chatbot RAG Dataset*. It was created to evaluate the model's retrieval capabilities in a RAG-based chatbot system. By incorporating elements such as message history, the new user message (which may or may not be a follow-up), and a question-answer format for the documents, it simulates a real-world retrieval scenario in a RAG chatbot system.

SynPerChatbotTopicsRetrieval This dataset is derived from the *Synthetic Persian Chatbot Dataset*. It was created to evaluate the model's retrieval capabilities for identifying the topics of conversations between the user and the chatbot.

SynPerChatbotRAGTopicsRetrieval This dataset is derived from the *Synthetic Persian Chatbot RAG Dataset*. It was created to evaluate the model's retrieval capabilities for identifying the topics of user-chatbot conversations, including conversations that have not necessarily been completed.

SynPerQARetrieval This dataset is derived from the *Synthetic Persian QA Dataset*. It is designed to evaluate the model's retrieval capabilities when the query is a question, and the documents are in the format of answers.

PersianWebDocumentRetrieval (Zinvandi et al., 2024) This dataset consists of queries collected from the Zarebin search engine and human-labeled documents, specifically curated for the task of information retrieval from the web.

NeuCLIR2023Retrieval (Lawrie et al., 2023) and NeuCLIR2022Retrieval (Lawrie et al., 2024) The NUECLIR dataset is designed for the Cross-Language Information Retrieval (CLIR) task, where systems process queries in one language (English) and retrieve relevant news articles written in a different language, such as Chinese, Persian, or Russian. The Persian collection within this dataset includes Persian documents and corresponding English queries, which are scheduled for release. The query object comprises both human-translated and machine-translated queries, facilitating monolingual retrieval and cross-language retrieval scenarios.

WikipediaRetrievalgMultilingual (Foundation, 2023) The dataset utilized in this study is derived from Cohere's Wikipedia-2023-11 dataset, which comprises a comprehensive collection of Wikipedia articles. Additionally, the dataset includes synthetically generated queries designed to facilitate information retrieval tasks.

MIRACL (Zhang et al., 2023) The MIRACL dataset is a multilingual dataset for information retrieval. It comprises documents from Wikipedia and human-generated queries corresponding to these documents. One of the languages supported by this dataset is Persian, which we utilize for the information retrieval task.

B.6 Reranking

MIRACLReranking MIRACL-Reranking is a subset of the MIRACL dataset, where, for each query, a collection of documents is retrieved and annotated as relevant or irrelevant.

WikipediaRerankingMultilingual (Foundation, 2023) The dataset is sourced from Cohere's Wikipedia-2023-11 dataset and includes synthetically generated queries.

B.7 Pair Classification

CExaPPC (Sadeghi et al., 2022) ExaPPC is an extensive paraphrase corpus consisting of monolingual Persian sentence-level paraphrases, derived from a variety of sources.

SynPerChatbotRAGFAQPC This dataset is derived from the *Synthetic Persian Chatbot RAG Dataset*. It evaluates the model's ability to identify relevant FAQ question-answer pairs corresponding to the user's new message in a chatbot conversation, considering the message history within a RAG-based system.

FarsiParaphraseDetection (Ghasemi, 2022) This dataset consists of 7,826 pairs of sentences in Persian, manually annotated for paraphrase detection.

FarsTail (Amirkhani et al., 2020) FarsTail dataset is a Persian textual entailment dataset de-

signed to facilitate natural language understanding tasks.

SynPerTextKeywordsPC This dataset is derived from the *Synthetic Persian Keywords Dataset*. It is designed to evaluate the model's ability to identify the keywords of a given text.

ParsinluEntail (Khashabi et al., 2021) The dataset focuses on a Persian textual entailment task, which involves determining whether one sentence entails another sentence. The questions are partially derived from translations of the SNLI dataset and partially generated by expert annotators.

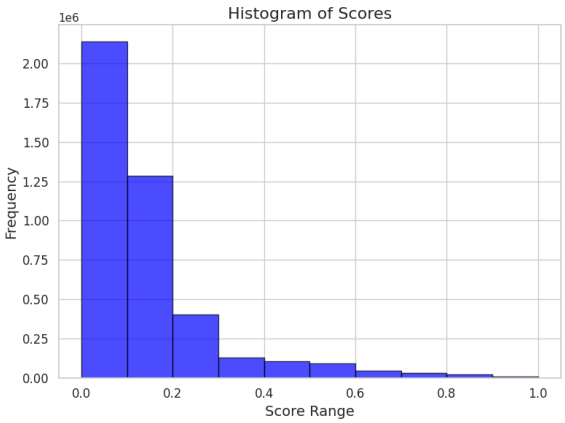
ParsinluQueryParaphPC (Khashabi et al., 2021) This study addresses a Persian query paraphrasing task, which involves determining whether two questions are paraphrases of each other. The dataset comprises questions that are partially generated using Google's auto-complete feature and partially translated from the Quora Paraphrasing dataset.

SynPerQAPC This dataset is derived from the *Synthetic Persian QA Dataset*. It is designed to evaluate the model's ability to identify the correct answers for given questions.

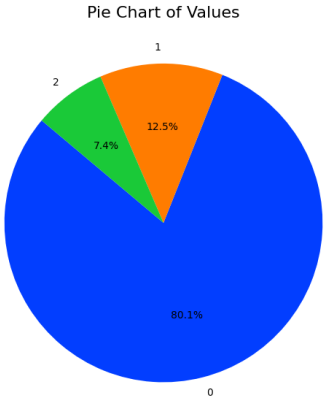
Figure 2: This figure illustrates the prompts used to generate the Synthetic Persian Chatbot Conversational Sentiment Analysis dataset. This dataset receives the chat subject, user tone, chatbot tone, and user emotion as input, corresponding to the placeholders "input subject", "input user tone", "input chatbot tone", and "emotion dictionary", respectively.

C Model Accuracy

The accuracy of each model, presented separately, is provided in Table 9.



(a)



(b)

Figure 3: Figure a shows the distribution of the Query to Query data based on similarity scores. Figure b illustrates the distribution of the Query to Query data according to the labels assigned, ranging from 0 to 2.

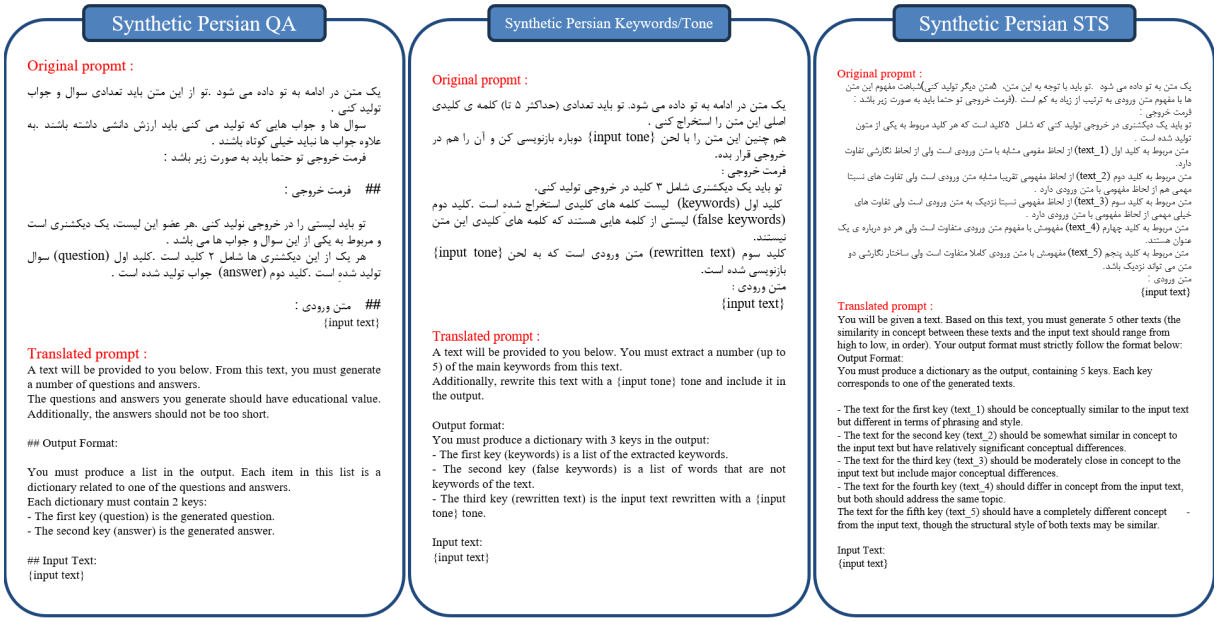


Figure 4: This figure illustrates the prompts used to generate the three datasets: a) Synthetic Persian QA, b) Synthetic Persian Keywords/Tone, and c) Synthetic Persian STS. The prompts are in Persian, with their translations included below each prompt. In the prompts, the placeholder "input text" is replaced with the text intended to generate the data. Additionally, in prompt b, there is an "input tone" placeholder to specify the desired tone.

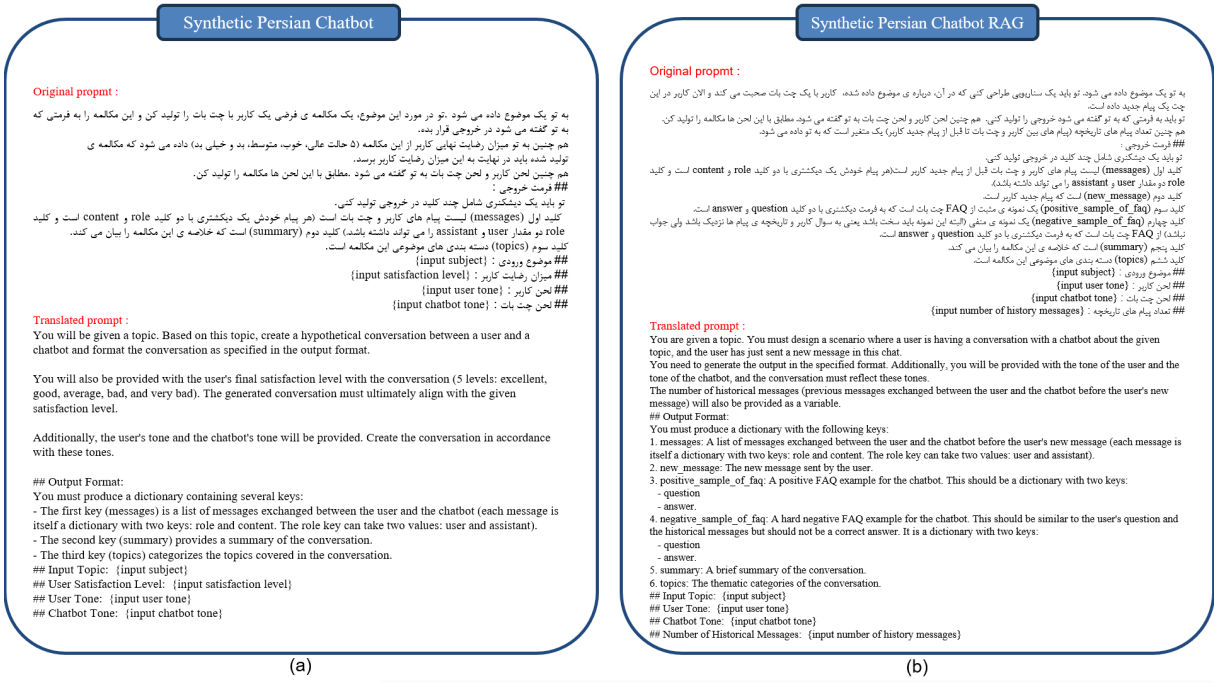


Figure 5: This figure illustrates the prompts used to generate the two datasets: a) Synthetic Persian Chatbot and b) Synthetic Persian Chatbot RAG. Both datasets receive the chat subject, user tone, and chatbot tone as input, corresponding to the placeholders "input subject", "input user tone", and "input chatbot tone", respectively. Additionally, in prompt a, the user's satisfaction level is provided through the placeholder "input satisfaction level", and in prompt b, the number of messages included in the chat history is specified using the placeholder "input number of history messages".

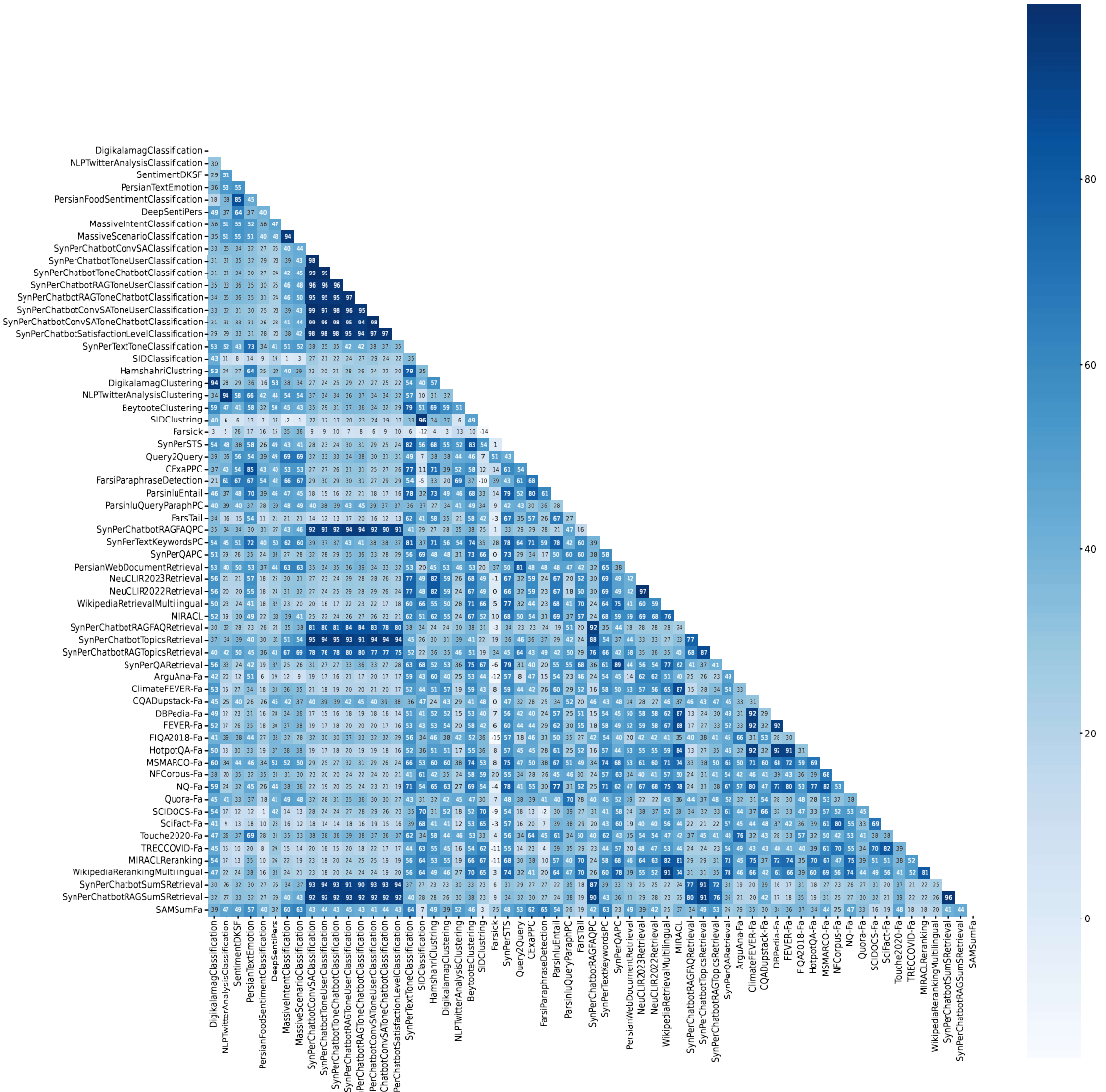


Figure 6: Similarity chart of different MTEB datasets.

Datasets	Type	Language	#Train	#Dev.	#Test
DigikalamagClassification	classification	Persian	6,896	767	852
NLPTwitterAnalysisClassification	classification	Persian	2,715	1,357	1,360
SentimentDKSF	classification	Persian	28,602	0	2,315
PersianTextEmotion	classification	Persian	5,558	0	1,390
PersianFoodSentimentClassification	classification	Persian	56,700	6,300	7,000
DeepSentiPers	classification	Persian	6,320	703	1,854
MassiveIntentClassification	classification	Multilingual	11,514	2,033	2,974
MassiveScenarioClassification	classification	Multilingual	11,514	2,033	2,974
SynPerChatbotConvSAClassification	classification	Persian	4,496	0	1,499
SynPerChatbotToneUserClassification	classification	Persian	8,709	0	1,537
SynPerChatbotToneChatbotClassification	classification	Persian	8,709	0	1,537
SynPerChatbotRAGToneUserClassification	classification	Persian	3,261	0	1,087
SynPerChatbotRAGToneChatbotClassification	classification	Persian	3,261	0	1,087
SynPerChatbotConvSAToneUserClassification	classification	Persian	4,496	0	1,499
SynPerChatbotConvSAToneChatbotClassification	classification	Persian	4,496	0	1,499
SynPerChatbotSatisfactionLevelClassification	classification	Persian	8,709	0	1,537
SynPerTextToneClassification	classification	Persian	16,587	0	2,928
SIDClassification	classification	Persian	8,712	0	3,735
HamshahriClustring	clustering	Persian	0	0	2,203
DigikalamagClustering	clustering	Persian	6,896	767	852
NLPTwitterAnalysisClustering	clustering	Persian	2,715	1,357	1,360
BeytooteClustering	clustering	Persian	0	0	95,851
SIDClustring	clustering	Persian	8,712	0	3,735
Farsick	STS	Persian	0	0	8,566
SynPerSTS	STS	Persian	70,155	0	12,385
Query2Query	STS	Persian	4,228,933	0	42,717
CExaPPC	pair classification	Persian	63,021	13,505	13,504
FarsiParaphraseDetection	pair classification	Persian	6,260	783	783
ParsinluEntail	pair classification	Persian	755	270	1,675
ParsinluQueryParaphPC	pair classification	Persian	1,830	898	1,916
FarsTail	pair classification	Persian	0	0	1,564
SynPerChatbotRAGFAQPC	pair classification	Persian	6,522	0	2,174
SynPerTextKeywordsPC	pair classification	Persian	33,174	0	5,856
SynPerQAPC	pair classification	Persian	500,106	0	55,568
PersianWebDocumentRetrieval	retrieval	Persian	245,692	0	175,472
NeuCLIR2023Retrieval	retrieval	Multilingual	0	0	2,258,678
NeuCLIR2022Retrieval	retrieval	Multilingual	0	0	2,266,190
WikipediaRetrievalMultilingual	retrieval	Multilingual	0	0	15,000
MIRACL	retrieval	Multilingual	0	0	2,213,743
SynPerChatbotRAGFAQRetrieval	retrieval	Persian	11,957	0	9,783
SynPerChatbotTopicsRetrieval	retrieval	Persian	34,084	0	11,128
SynPerChatbotRAGTopicsRetrieval	retrieval	Persian	14,590	0	7,648
SynPerQARetrieval	retrieval	Persian	520,695	0	298,426
ArguAna-Fa	retrieval	Persian	0	0	10,080
ClimateFEVER-Fa	retrieval	Persian	0	0	5,421,274
CQADupstack-Fa	retrieval	Persian	0	0	480,902
DBPedia-Fa	retrieval	Persian	0	0	4,651,208
FEVER-Fa	retrieval	Persian	5,556,643	0	5,424,495
FIQA2018-Fa	retrieval	Persian	71,804	0	59,344
HotpotQA-Fa	retrieval	Persian	5,403,329	0	5,248,139
MSMARCO-Fa	retrieval	Persian	9,374,574	0	8,845,925
NFCorpus-Fa	retrieval	Persian	114,208	0	15,967
NQ-Fa	retrieval	Persian	0	0	2,685,669
Quora-Fa	retrieval	Persian	0	0	538,606
SCIDOCS-Fa	retrieval	Persian	0	0	30,585
SciFact-Fa	retrieval	Persian	6,102	0	5,522
Touche2020-Fa	retrieval	Persian	0	0	383,477
TRECCOVID-Fa	retrieval	Persian	0	0	196,005
MIRACLReranking	reranking	Multilingual	0	0	1,314
WikipediaRerankingMultilingual	reranking	Multilingual	0	0	1,500
SynPerChatbotSumSRetrieval	summary retrieval	Persian	8,709	0	1,537
SynPerChatbotRAGSumSRetrieval	summary retrieval	Persian	3,261	0	1,087
SAMSumFa	summary retrieval	Persian	14,045	0	1,561

Table 7: The number of documents in each of the FaMTEB datasets.

car-news	news	computer	attire	health	fun	cookery	art	iran	psychology	baby	sport	wedlock	religious	scientific	mode	housekeeping	pictures	job
12669	12000	8603	6109	5333	5241	4869	4741	4441	4239	4107	3889	3856	3700	3578	3047	2951	1251	1227

Table 8: The number of samples in each category of dataset BeytooteClustering.

Figure 7: Example of classification datasets.

	sentence-transformer-parsbert-fa	RoBERTa-WLNI	BERT-WLNI	faBert	ParsBERT	paraphrase-multilingual-MiniLM-L12-v2	LaBSE	TookaBERT-Base	Tooka-SBERT	GTE-multilingual-base	multilingual-e5-base	multilingual-e5-large	BGE-m3-unsupervised	BGE-m3	Jina-embeddings-v3
Classification															
PersianFoodSentimentClassification	64.25	64.52	64.17	70.3	73.75	73.46	72.09	72.01	80.05	77.49	75.05	76.31	77.34	83.4	83.57
SynPerChatbotConvSAClassification	62.17	59.44	58.23	70.51	77.1	57.51	75.41	78.07	76.38	63.29	64.61	60.77	63.15	61.03	71.57
SynPerChatbotConvSAToneChatbotClassification	70.01	68	65.39	88.43	91.21	56.6	66.77	88.97	60.75	49.09	63.18	58.07	54.42	50.55	51.88
SynPerChatbotConvSAToneUserClassification	55.72	62.01	61.77	48.36	72.68	50.35	53.63	69.6	56.46	51.86	48.85	52.6	51.3	48.85	52.86
SynPerChatbotSatisfactionLevelClassification	26.14	21.98	23.05	31.87	35.6	22.04	35.02	37.22	37.18	30.82	25.32	25.23	26.04	24.72	35.43
SynPerChatbotRAGToneChatbotClassification	38.65	43.02	42.41	58.41	61.39	38.41	42.97	58.31	38.69	32.41	35.15	37.16	35.16	35.45	33.16
SynPerChatbotRAGToneUserClassification	47.07	58.03	53.84	50.9	62.37	44.68	54	62.59	51.32	48.45	44.9	50.73	50.7	48.47	49
SynPerChatbotToneChatbotClassification	48.73	48.99	48.26	67.53	74.99	41.14	52.42	72.46	47.68	33.64	42.36	41.5	42.38	37.92	36.23
SynPerChatbotToneUserClassification	44.93	56.96	52.97	41.76	60.17	41.67	51.15	57.25	46.91	43.2	39.98	46.73	45.75	42.71	45.18
SynPerTextToneClassification	55.86	71.12	73.12	91.3	89.72	46.39	58.71	89.78	51.8	51.7	63.69	70.19	61.54	55.67	50.69
SIDClassification	55	50.41	48.26	54	55.7	54.46	56.71	58.75	53.24	60.62	60.73	61.37	60.32	59.62	61.68
DeepSentiPers	42.51	43.38	44.28	42.41	50.63	55.91	60.91	54.53	64.43	57.95	61.9	60.95	58.89	67.51	65.26
PersianTextEmotion	38.81	37.59	37.5	48.8	48.18	45.37	53.45	53.26	57.01	51.5	54.85	61.88	58.73	61.13	51.88
SentimentDKSF	49.17	49.87	51.52	59.04	59.69	65.62	67.67	58.16	69.57	67.52	71.26	71.07	66.34	75.35	75.15
NLPTwitterAnalysisClassification	73.68	70.54	70.74	70.29	70.71	74.98	74.93	70.28	74.72	75.62	74.67	75.99	77.75	76.93	75.97
DigikalamagClassification	79.74	74.79	74.58	82.59	81.8	74.66	85.12	82.37	72.95	82.93	86.78	87.05	86.31	86.03	84.88
MassiveIntentClassification (fa)	44.19	51.43	52.78	55.09	60.07	61.03	62.33	59.98	63.73	62.29	59.51	63.74	66.17	69.44	72.6
MassiveScenarioClassification (fa)	51.78	59.53	58.24	58.46	62.6	65.89	67.43	64.11	67.45	67.88	63.92	67.55	72.37	73.29	81.78
Clustering															
BeytooteClustering	61.95	61.92	60.75	53.81	51.6	63	56.12	55.27	55.62	62.52	59.16	61.5	61.44	60.71	63.4
DigikalamagClustering	60.53	46.24	55.82	38.71	50.73	48.69	44.07	45.55	41.03	34.39	38.63	39.89	47.48	39.56	43.3
HamshahriClustring	75.57	68.85	67.37	65.08	64.84	63.84	67.09	66.43	63.28	69.83	67.83	67.42	67.55	69.48	66.88
NLPTwitterAnalysisClustering	79.18	76.24	77.85	74.55	74.46	78.97	76.12	70.06	82.24	80.82	78.18	78.48	80.63	80.9	80.69
SIDClustring	46.91	39.8	39.79	43.77	39.12	36.11	39.4	41.3	40.08	38.86	38.79	38.65	41.02	38	41.5
Pair Classification															
FarsTail	58.92	54.98	56.09	57.15	57.79	64.84	62.93	55.6	81.52	72.65	70.76	69.74	69.77	73.14	71.85
CExaPPC	82.72	93.7	94.29	90.88	91.13	95.95	98.97	94.55	98.8	98.42	98.7	98.97	97.26	99.09	97.41
SynPerChatbotRAGFAQPC	54.02	55.81	52.54	58.26	50.84	60.75	62.68	60.32	71.32	66.77	65.42	62.9	65.03	64.43	62.06
FarsiParaphraseDetection	90.38	92.02	93.01	77.99	93.69	96.82	94.34	86.98	95.99	97.06	96.39	97.57	96.86	95.57	96.6
SynPerTextKeywordsPC	86.94	82.59	81.97	76.02	71.25	89.93	87.88	77.68	94.93	96.4	95.73	94.79	96.21	97.04	94.96
SynPerQAFaPC	65.13	73.11	73.52	66.6	67.84	80.6	83.07	66.2	86.34	91.66	94.24	95.16	93.2	93.76	93.38
ParsinluEntail	55.14	55.9	55.58	57.28	57.79	69.19	60.76	56.39	77.7	66.55	64.81	65.43	59.81	68.65	65.82
ParsinluQueryParaphPC	71.18	60.72	61.57	65.67	62.8	81.15	80.4	67.83	89.7	87.09	86.3	87.75	86.41	89.98	87.63
Reranking															
MIRACLReranking (fa)	18.34	17.38	18.17	23.72	18.9	30.83	29.05	14.4	35.87	55.05	57.36	59.36	48.78	60.92	42.91
WikipediaRerankingMultilingual (fa)	61.47	72.11	73.28	77.43	73.86	80.8	82.42	73.95	80.71	84.38	86.78	89.32	90.71	88.21	79.61
Retrieval															
SynPerQARetrieval	20.36	26.59	25.02	42.94	24.95	52.45	53.99	26.88	65.02	77.44	85.59	87.35	85.14	86.27	85.4
SynPerChatbotTopicsRetrieval	2.38	3.52	4.23	0.05	0.15	12.28	6.2	0.16	10.76	28.07	15.37	11.82	10.59	19.18	18.75
SynPerChatbotRAGTopicsRetrieval	4.33	4.59	4.72	0.09	1.19	16.39	12.1	0.45	18.93	30.97	20.11	19.24	13.22	19.91	24.26
SynPerChatbotRAGFAQRetrieval	6.76	7.46	6.37	10.02	5.09	19.22	18.82	12.24	24.3	31.47	28.49	23.48	30.84	32.04	47.46
PersianWebDocumentRetrieval	12.61	8.14	12.55	7.95	10.04	14.31	28.21	10.85	43.9	44.15	46.72	46.76	38.18	44.09	40.32
NeuCLIR2022Retrieval	1.33	3.9	3.23	2.92	0.38	19.78	2.56	0.02	26.96	36.67	9.75	5.3	12.12	15.48	18.25
NeuCLIR2023Retrieval	6.6	5.27	5.02	12.1	1.86	26.34	21.52	4.63	36.47	50.93	46.1	46.67	46.53	52.2	51.45
WikipediaRetrievalMultilingual (fa)	35.63	37.34	41.29	63.67	48.41	62.15	67.06	46.14	79.02	84.94	88.11	90.4	91.19	89.32	89.02
MIRACLRetrieval (fa)	1.95	4.34	4.35	8.24	4.52	13.33	10.53	2.21	21.32	53.89	57.48	59.01	39.93	60.9	55.21
ClimateFEVER-Fa	1.13	2.68	2.15	5.06	2.05	12.23	3.73	0.39	9.47	18.83	12.6	12.75	16.41	24.31	29.87
FEVER-Fa	0.7	1.11	1.42	1.7	0.59	18	7	0.41	8.44	61.33	48.05	41.56	44.74	55.99	63.75
DBPedia-Fa	1.92	1.13	2.19	2.87	1.92	11.53	10.78	1.2	13.77	29.2	28.74	30.36	22.47	29.85	31.84
HotpotQA-Fa	0.22	0.85	0.85	6.52	3.37	12.39	11.94	2.33	16.44	49.04	55.33	60.15	39.24	56.54	51.43
MSMARCO-Fa	1.04	1.5	1.34	2.02	1.14	7.89	6.43	1.23	9.33	23.33	26.88	30.92	21.38	29.09	29.85
NQ-Fa	0.62	1.88	1.75	3.49	1.27	11.49	7.88	1.08	11.97	38.8	39.84	44.82	26.69	46.62	50.33
ArguAna-Fa	20.59	18.77	14.94	22.24	21.8	36.45	36.13	27.51	31.88	50.4	43.19	45.5	55.28	50.4	34.88
CQADupstackRetrieval-Fa	2.46	4.89	4.45	9.46	4.49	18.35	16.65	3.09	18.09	26.03	29.87	31.59	32.45	31.72	27.91
FiQA2018-Fa	0.87	2.61	2.22	3.1	1.58	10.31	6.35	1.82	11.22	26.47	23.17	30.15	27.39	30.38	34.43
NFCorpus-Fa	5.36	3.45	4	7.14	3.36	14.83	15.52	4.86	19.7	25.61	25.47	28.59	28.28	29.47	28.21
QuoraRetrieval-Fa	47.21	46.38	47.38	52.45	49.29	73.61	72.52	51.72	76.67	77.99	77.26	79.96	80.14	82.18	59.44
SCIDOCS-Fa	1.85	2.94	1.66	3.74	1.56	9.28	5.8	1.97	9.07	12.72	11.76	11.58	13.55	14.56	14.52
SciFact-Fa	5.95	6.48	7.75	18.86	10.33	31.54	34.57	13.54	37.19	56.15	57.79	59.69	57.76	60.52	61.51
Touche2020-Fa	1.44	7.04	5.46	2.14	1.14	16.47	4.59	1.42	13.22	24.69	22.48	26.19	16.15	22.87	26
STS															
Farsick	50.16	48.45	50.62	49.27	53.57	66.72	64.42	52.04	69.64	70.95	69.93	70.67	67.29	71.75	76.96
SynPerSTS	67.67	72.66	71.99	77.16	73.98	83.31	88.47	74.55	89.08	86.89	86.7	87.98	86.91	87.59	88.73
Query2Query	47.38	43.66	46.21	29.61	55.37	51.69	66.28	59.08	70.52	69.41	66.71	67.49	69.22	69.71	70.27
Summary Retrieval															
SAMSumFa	26.2	5.49	8.65	14.45	11.23	45.04	83.74	9.31	88.19	85.5	92.97	92.42	92.47	97.88	96.89
SynPerChatbotSumSRetrieval	3.21	1.03	1.95	3.18	1.36	17.37	18.02	3.2	32.73	36.78	25.09	27.6	36.61	32.13	36.99
SynPerChatbotRAGSumSRetrieval	14.71	10.62	8.45	10.34	8.68	33.32	40.76	12.35	56.27	60.37	45.68	49.81	55.93	53.21	62.63

Table 9: The accuracy of each model on all datasets, presented separately.

Figure 8: Example of clustering datasets.

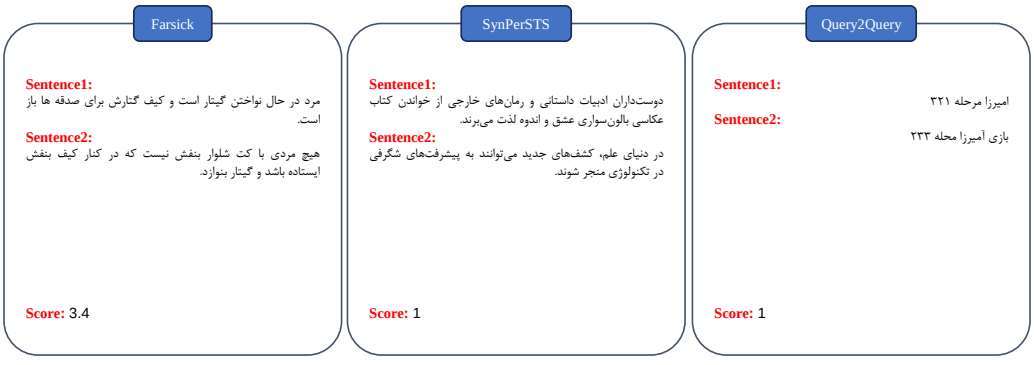


Figure 9: Example of STS datasets.

Figure 10: Example of retrieval datasets.

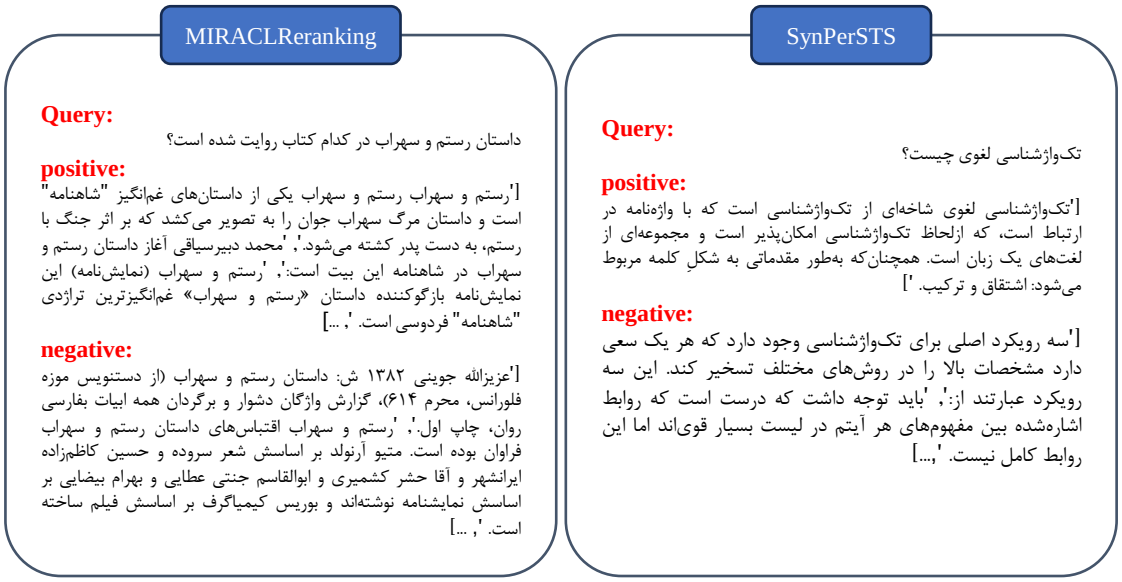


Figure 11: Example of reranking datasets.

Figure 12: Example of pair classification datasets.

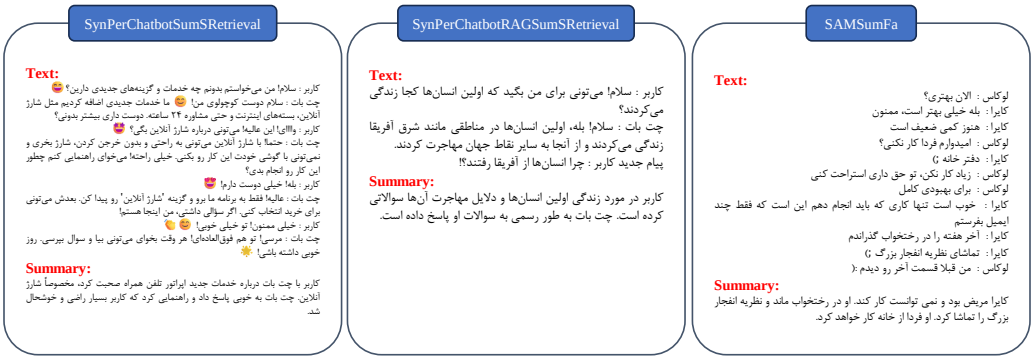


Figure 13: Example of summary retrieval datasets excluding BEIR-Fa datasets.