Advancing Retrieval-Augmented Generation for Persian: Development of Language Models, Comprehensive Benchmarks, and Best Practices for Optimization

Sara Bourbour Hosseinbeigi[1], Mohammad Hossein Shalchian[2], Mohammad Ali Seif Kashani[3], Sina Asghari[4], Mohammad Amin Abbasi[5]

[1] Department of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
s.bourbour@modares.ac.ir

[2] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
mo.shalchian@sharif.edu

[3] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
ma.seifkashani@ce.sharif.edu

[4] Department of Computer Science, Iran University of Science and Technology, Tehran, Iran
sina_asghari@mathdep.iust.ac.ir

[5] Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
m_abbasi1378@comp.iust.ac.ir

Abstract

This paper examines the specific obstacles of constructing Retrieval-Augmented Generation (RAG) systems in low-resource languages, with a focus on Persian's complicated morphology and versatile syntax. The research aims to improve retrieval and generation accuracy by introducing Persian-specific models, namely MatinaRoberta (a masked language model) and MatinaSRoberta (a fine-tuned Sentence-BERT), along with a comprehensive benchmarking framework. Three datasets—general knowledge (PQuad), scientifically specialized texts, and organizational reports—were used to assess these models after they were trained on a varied corpus of 73.11 billion Persian tokens. The methodology involved extensive pretraining, fine-tuning with tailored loss functions, and systematic evaluations using both traditional metrics and the Retrieval-Augmented Generation Assessment (RAGAS) framework. The results show that MatinaSRoberta outperformed previous embeddings, achieving superior contextual relevance and retrieval accuracy across datasets. Temperature tweaking, chunk size modifications, and document summary indexing were explored to enhance RAG setups. Larger models like Llama-
3.1 (70B) consistently demonstrated the highest generation accuracy, while smaller models faced challenges with domain-specific and formal contexts. The findings underscore the potential for developing RAG systems in Persian through customized embeddings and retrieval-generation settings and highlight the enhancement of NLP applications such as search engines and legal document analysis in low-resource languages.

Keywords: Retrieval-Augmented Generation, Large Language Models, Benchmarking, Persian, Sentence Embeddings

1. Introduction

In natural language processing (NLP) (Chowdhary, 2020), retrieval-augmented generation (RAG) (Gao et al., 2023) systems enhance generative outputs by incorporating external knowledge bases, addressing the limitations of standalone large language models (LLMs) (Minaee et al., 2024) which often suffer from outdated or insufficient knowledge. Recent advancements highlight the potential of integrating high-quality retrieval mechanisms with fine-tuned language models to reduce hallucinations (Tonmoy et al., 2024) and improve factual consistency. However, applying these techniques to languages with limited resources, such as Persian, presents unique challenges. Persian's rich morphology, flexible syntax, and scarcity of annotated resources necessitate specialized approaches to both retrieval and language modeling.

This project addresses these challenges by developing Persian-specific models, including a Sentence-BERT (Reimers & Gurevych, 2019) and a masked language model, aimed at optimizing retrieval processes in RAG systems. A robust framework is established to systematically assess their performance across formal, technical, and general knowledge settings. The study also examines how various retrieval and generation configurations—such as document chunking, temperature settings, and summary-based indexing—impact key performance metrics like generative faithfulness, contextual relevance, and retrieval accuracy.

By achieving these objectives, this research advances LLMs in low-resource contexts, providing insights transferable to other underrepresented languages. It focuses

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7320–7330
11–16 May 2026.

on three representative datasets for evaluating RAG systems in Persian: (1) the PQuad dataset (Darvishi et al., 2023) for general knowledge tasks, (2) a scientific-specialized dataset for technical content, and (3) formal organizational reports, each reflecting diverse linguistic challenges and contextual needs. The study evaluates the performance of Persian-specific masked language models and Sentence-BERT models alongside other state-of-the-art LLMs.

This research makes the following contributions:
1. Development of Persian-specific masked language and Sentence-BERT models, filling crucial gaps in NLP resources. 2. Establishment of a comprehensive benchmark for systematic evaluation of RAG systems in low-resource languages. 3. Identification of best practices for RAG optimization, offering practical insights into model configurations and trade-offs between retrieval and generation.

These contributions enhance the understanding of RAG systems, opening the door for improved applications in search engines, legal document analysis, and domain-specific retrieval systems by tailoring solutions to the linguistic features of Persian.

The paper is organized as follows: Section 2 reviews related work on retrieval-augmented generation systems and the application of LLMs to low-resource languages. Section 3 describes the development of MatinaRoberta, a Persian-specific masked language model, followed by Section 4, which introduces MatinaSentenceRoberta, a fine-tuned Sentence-BERT model for Persian retrieval tasks. Section 5 outlines the benchmarking methodology, datasets, and evaluation metrics used to assess RAG performance. Section 6 presents the results and discussion, analyzing model performance across diverse tasks and configurations. Finally, Section 7 concludes with a summary of contributions, implications for NLP in low-resource settings, and future research directions.

2. Related Work

2.1 Searching for Best Practices in Retrieval-Augmented Generation

(Wang et al., 2024) investigate optimization in RAG systems via query-dependent retrieval, examining components like query classification, document retrieval, reranking, and summarization to boost efficiency and accuracy. They focus on multimodal retrieval (text-to-image and image-to-text) for visual question-and-answer tasks, proposing a balance between performance and complexity through systematic comparisons. This research advances RAG systems by offering insights that are applicable across various domains, including multimodal and text-only environments.

2.2 Benchmarking Large Language Models in Retrieval-Augmented Generation

(Chen et al., 2024) assess large language models' performance in RAG tasks, analyzing challenges linked to integrating external knowledge. They develop the Retrieval-Augmented Generation Benchmark (RGB) to evaluate noise robustness, negative rejection, information integration, and counterfactual robustness. The study reveals that while LLMs demonstrate some noise resilience, they struggle with rejecting incorrect data, integrating information, and managing counterfactuals. This contributes to improving RAG systems by highlighting inefficiencies and providing a new evaluation benchmark. It emphasizes retrieval's potential to enhance LLM performance by reducing hallucinations and updating outdated knowledge, yet identifies challenges in handling misinformation and complex integrations. The study offers significant insights for future research aiming to boost RAG system reliability and robustness.

2.3 CRUD-RAG: A Chinese Benchmark for RAG Systems

(Lyu et al., 2024) present a benchmark for RAG systems focusing on Chinese language tasks, dividing RAG operations into CRUD (Create, Read, Update, Delete), with specific evaluation tasks like text continuation, question answering, multi-document summary, and modifying hallucinations. Their work addresses limitations of existing RAG evaluations, typically centered on question-answering, by considering diverse components such as retrievers, chunk sizes, and embedding models. The research enriches RAG system optimization across varied tasks and scenarios and sets a new benchmark for Persian NLP within low-resource languages by focusing on structural and morphological challenges. This framework has broader implications for RAG systems in similar languages.

3. MatinaRoBERTa

Effective Retrieval-Augmented Generation (RAG) systems rely heavily on quality sentence embeddings for precise retrieval and generation. Sentence-BERT is vital in this pipeline as it provides dense vector representations that encapsulate semantic similarities, crucial for accurate RAG retrieval. However, its efficacy depends on a robust

language model pretrained on diverse, high-quality data.

Therefore, we continuously pretrained XLM-RoBERTa (Conneau et al., 2019) Large on 73.11 billion Persian tokens. The approach used MLM for tuning into a Sentence-BERT framework, excluding the Next Sentence Prediction (NSP) task, aligning with research showing better performance without NSP. This method captures the complex morphology and syntax of Persian, establishing a solid base for creating high-quality embeddings for RAG systems.

3.1 Corpus Details

The pretraining corpus (Bourbour Hosseinbeigi et all., 2025) was curated from diverse Persian text sources, ensuring wide-linguistic coverage as outlined in Table 1. Scientific articles provided
3.55 billion tokens, adding formal, structured content. Books on general, educational, and religious topics contributed 2.84 billion tokens, spanning multiple genres. Social media delivered
2.34 billion tokens of informal language, while Persian websites offered 14.78 billion tokens across news, blogs, and forums. Additionally,
49.6 billion tokens from Common Crawl datasets were included to capture modern language use. These sources collectively furnished an extensive range of Persian text types, aiding effective model generalization. See **Table 1**.

Dataset	**# Tokens**
Scientific articles	3.55B
Books	2.84B
Social Media	2.34B
Persian Websites	14.78B
Common Crawl	49.6B
Sum	**73.11B**

Table 1 Token Distribution Across Pretraining Datasets.

3.2 Preprocessing

Before training, the data underwent comprehensive preprocessing, including deduplication and noise removal, which boosted the corpus's integrity and accuracy, especially important for Persian as a low-resource language.

3.3 Training Procedure

The training process was streamlined for efficiency over a large dataset. The MLM objective required predicting masked tokens via surrounding context, adeptly handling Persian's intricate language features. Over one week, training deployed eight NVIDIA A800 GPUs with DeepSpeed's Stage 0 (Rasley et al., 2020) for enhanced performance. The process utilized a maximum sequence length of 512 tokens and FP16 mixed precision for streamlined computation. A learning rate of 5e-5 with a linear warm-up schedule was chosen. The batch size was 30 per device, with two gradient accumulation steps, making an effective batch size of 480; the evaluation batch size was set at eight. Optimization used the AdamW optimizer with betas (0.9, 0.999) and epsilon of 1e-8, managed by a linear scheduler. Training lasted one epoch, fully exposing the model to the dataset.

Model	Arman Emo	Pars-ABSA	Taghche	Snapp Food	Digikala	Digimag	Persian News	PQUAD	Reading Comprehension	DeepSentPars	PEYMA
MatinaRoberta	**56.54**	**74.92**	**59.34**	**88.7**	**72.59**	**96.37**	**98.05**	86.82	**56.82**	**83.21**	85.65
TookaBERT	52.87	74.65	58.17	87.26	66.56	93.91	96.95	86.73	44.89	82.96	86.09
AriaBERT	38.23	74.59	58.58	87.93	67.06	92.38	97.63	83.14	37.98	73.21	35.78
XLM-RoBERTa	32.48	74.18	54.59	86.28	63.06	91.58	96.9	**87.6**	42.55	73.39	**87.94**
mBERT	6.74	68.15	54.31	86.18	61.06	69.69	90.17	85.94	49.63	55.78	65.32

Table 2 Results of Masked Language Models Evaluation.

3.4 Applications and Impact

MatinaRoberta excels in various Persian NLP tasks, such as text classification, question answering, semantic search, and generating contextual embeddings. It's especially useful in RAG systems, offering precise retrieval and grounded generation. This significantly enhances applications like Persian-language search engines, knowledge assistants, and domain-specific retrieval, advancing performance for low-resource languages. We compared our model with TookaBERT (SadraeiJavaheri et al., 2024), AriaBERT (Ghafouri et al., 2023), XLM-RoBERTa, and mBERT (Pires et al., 2019). Evaluation results compared with other models are detailed in Table 2.

4. MatinaSRoBERTa

Post-masked language modeling, the model underwent fine-tuning to enhance its role in RAG systems, specifically as a similarity-based text embedding model. This process aimed to align embeddings for semantically related text pairs and differentiate dissimilar pairs, crucial for improving retrieval accuracy and generation quality in RAG pipelines. Diverse datasets

spanning linguistic and semantic tasks, such as question-answer matching, entailment classification, paraphrase detection, and triplet-based semantic similarity, were used. Such efforts refined the model's capability to retrieve and fuse relevant information, allowing for the generation of contextually accurate, semantically rich responses, establishing it as a leading solution in RAG frameworks.

Fine-tuning incorporated key datasets like the multilingual Miracle Project's triplets in English, Arabic, and Persian, as well as the ParsiNLU (Khashabi et al., 2021) datasets for question pair similarity and entailment classification. Persian domain-specific datasets enriched the model's understanding of specialized settings, including QA pairs from PQuad, paraphrase pairs, Wikipedia-derived triplets, and domain-focused texts in religion, law, education, and tourism. Additionally, Persian instructional datasets and bilingual Persian-English translation pairs improved the model's ability to manage diverse language constructs and semantic relationships.

The procedure utilized loss functions designed for the datasets' structures and tasks. Multiple Negatives Ranking Loss was applied to anchor-positive pairs to leverage other samples in a batch as negatives, optimizing semantic similarity learning. Contrastive Loss was used for binary semantic similarity classification, while Softmax Loss facilitated tasks requiring three-class entailment classification. Triplet Loss was applied to datasets containing anchor, positive, and negative samples, ensuring embeddings for anchors and positives surpassed those for anchors and negatives. These loss functions

enabled the model to adeptly tackle a broad spectrum of semantic tasks.

Embeddings generated were dense 1024-dimensional vectors, achieved by averaging token embeddings from transformer layers. Inspired by Sentence-BERT, the approach yielded efficient, high-quality semantic representations. Training spanned 3 hours on 1 NVIDIA A800 GPU. Training parameters, detailed in Table 3, supported a structured and efficient regimen. See **Table 3**.

Config	Value
Train batch_size	30
Gradient accumulation steps	2
Weight decay	0.01
Num train epochs	4
Lr scheduler type	polynomial
Warmup ratio	0.4
FP16	True

Table 3 Token Distribution Across Pretraining Datasets.

Config	Value
max tokens	2048
chunk size	1024
chunk overlap	256
similarity top k	5
temperature	0.25

Table 4 The base parameters used for evaluations.

5. Benchmarking RAG

This section describes how we evaluated the performance of Retrieval-Augmented Generation (RAG) in Persian, testing various embedding and large language models across three datasets: general knowledge, scientific-specialized texts, and formal organizational documents. This study examines strategies for optimizing RAG systems within the Persian language context, addressing linguistic challenges and model adaptability to different content types.

5.1 Evaluation Datasets

To thoroughly evaluate the retrieval and generation capabilities of the models, three distinct datasets were employed, each reflecting a different type of textual content and linguistic complexity.

5.1.1 General Knowledge Dataset

Derived from PQuad, a Persian-language dataset with Wikipedia-based questions, this collection includes 80,000 questions, 25% of which are unanswerable, split into training (63,994), validation (7,976), and test (8,002) sets. Its diverse Wikipedia topics evaluate RAG models' ability to generalize across varied Persian text contexts and features.

5.1.2 Scientific-Specialized Dataset

Constructed from content of the Persian textbook General Physical Education (Mazyar & Azaryan, 2018), this dataset evaluates model performance in handling domain-specific language. By cleaning and creating multiple-choice questions using GPT-4, this dataset tests LLMs' retrieval and processing of specialized terms crucial in technical or academic fields.

5.1.3 Organizational Report Dataset

Utilized the Fundamental Transformation Document of Education in Iran, showcasing formal policy document language with socio-political terminology. This dataset tests model abilities to handle formal texts, with MCQs generated via GPT-4o assessing the processing of policy-oriented and culturally intricate documents.

5.2 Embedding Model Evaluation

The We assessed various embedding models including MatinaSRoberta, LaBSE (Feng et al., 2022), L12-V2, Qwen2-7 (Yang et al., 2024), and Alibaba/gte. Selected for their efficacy in low-resource languages like Persian, these models produce high-quality sentence embeddings and were tested for retrieval relevance across the three datasets, focusing on accuracy in limited-resource and complex linguistic settings.

5.3 Experimental Setup

5.3.1 Baseline Evaluation of Large Language Models

Using the LlamaIndex framework, LLMs like LLaMA 3.1 (8B & 70B) (Dubey et al., 2024), Qwen 2 (7B & 72B), Gemma 1.1 (Mesnard et al., 2024), and Gemma 2 (Riviere et al., 2024) were evaluated through Multiple-Choice Question Answering (MCQA) tasks on all three datasets. These tests established a baseline for understanding model performance across different Persian document types and complexity. See **Table 4**.

5.3.2 Temperature Tuning

Temperature tuning was utilized to optimize LLaMA 3.1 8B's output variety, testing four settings (0, 0.25, 0.5, and 0.75) to balance accuracy and diversity in Persian text responses, taking into account its flexible structure and polysemy.

Model	PQUAD				Scientific-Specialized				Organizational Report			
	1th	2th	3th	Avg	1th	2th	3th	Avg	1th	2th	3th	Avg
MatinaSRoberta	8231	475	142	**47.70**	766	206	22	**42.23**	356	150	32	21.02
Ahd	7231	604	181	42.70	715	206	76	40.70	412	106	52	**22.52**
LaBSE	4713	1041	457	30.85	403	172	91	25.41	412	106	52	22.46
L12-V2	4087	843	413	26.56	541	111	46	29.23	198	95	30	12.22
Qwen2-7	3979	608	278	24.85	87	83	83	7.88	46	51	85	4.879
Alibaba/gte-large	1341	476	313	9.78	290	121	107	18.84	310	67	120	17.77
Alibaba/gte	814	359	256	6.32	298	43	58	16.04	174	66	46	10.51

Table 5 Performance of Sentence-Transformer Models Across Datasets.

5.3.3 Chunk Size Testing

We tested chunk sizes of 512, 1024, and 2048 tokens using the LLaMA 3.1 8B model to evaluate retrieval performance. Smaller chunks improved precision but increased computational demand, whereas larger chunks increased efficiency but reduced precision in detail-sensitive texts.

5.4 RAGAS Framework

In this study, we used the Retrieval-Augmented Generation Assessment (RAGAS) framework (Shahul et al., 2024) to evaluate the performance of various Retrieval-Augmented Generation models. It provides a structured, automated evaluation across multiple metrics, offering a comprehensive assessment of both retrieval and generation capabilities. RAGAS

was chosen for its flexibility in evaluating models without reliance on ground truth data, which is scarce for low-resource languages like Persian. Traditional benchmarks often lack for these languages, making a flexible assessment like RAGAS necessary.

While the MCQA framework focused on generative capabilities, RAGAS provided a more comprehensive evaluation of both retrieval and generation. Key metrics in RAGAS include factual consistency, context relevance, retrieval accuracy, and model precision when handling queries. By using both MCQA and RAGAS, we ensured a broad range of tasks were covered—from structured multiple-choice tests to general retrieval and generation performance. This approach allowed for assessing models' effectiveness in practical scenarios, where factual accuracy and contextual relevance are crucial.

5.5 Evaluation Metrics

5.5.1 Evaluation Metrics for Embedding Models

To gauge the performance of each embedding model, the following metrics were used:

• **First Result Accuracy:** Percentage of queries with the correct answer as the first result.

• **Second Result Accuracy:** Percentage where the correct answer appeared second.

• **Third Result Accuracy:** Percentage with the correct answer as the third result.

• **Total Retrievals:** Total number of correct answers across all results.

• **Overall Score:** A composite metric factoring in correct retrievals for a comprehensive view of each model's performance.

The Overall Score was calculated using a weighted system: First Result accuracy was given a factor of 3, Second Result accuracy a factor of 2, and Third Result accuracy a factor of 1. These metrics provided insights into how well the embedding models retrieved relevant information across various text types, reflecting both generalization and specificity.

5.5.2 Evaluation Metrics for Large Language Models

Large language models were evaluated with the RAGAS framework and an additional metric for Multiple-Choice Question Answering:

• **Accuracy:** Assesses LLMs' performance in the Multiple-Choice Question Answering task via the percentage of correct answers. This metric indicates LLMs' ability to comprehend and generate correct answers using retrieved information, serving as a key performance indicator.

• **Faithfulness:** Evaluates factual consistency between the generated answer and the retrieved context, with scores from 0 to 1. Higher values reflect better factual accuracy.

• **Answer Relevance:** Measures how well the answer addresses the prompt. Higher scores are given for complete, concise answers free from redundancy, while lower scores indicate incomplete or irrelevant responses, based on the question, retrieved context, and generated answer.

• **Context Precision:** Determines if the retrieved context correctly prioritizes relevant items, ranking them higher. Scores range from 0 to 1, with higher scores indicating better ranking effectiveness.

• **Context Recall:** Assesses retrieval support for the ground truth answer. Scores range from 0 to 1, with higher values indicating stronger performance. A reference-free version, context utilization, is used when ground-truth answers aren't explicitly provided.

These metrics offer a comprehensive view of the LLMs' abilities to generate accurate, relevant answers and retrieve supportive context, ensuring quality of answers and factual consistency.

6. Results and Discussion

6.1 Embedding Models Evaluation

The embedding models' evaluation focused on their ability to retrieve relevant information. Results, including retrieval accuracy, are shown in Table 5. Retrieval accuracy was assessed by the rank of relevant results. MatinaSRoberta excelled across all datasets, particularly PQuad and scientific-specialized datasets, surpassing LaBSE and L12-V2 by efficiently handling Persian's linguistic features.

MatinaSRoberta's success with the PQuad dataset is attributed to its adept handling of Persian's flexible word order and morphology. Conversely, LaBSE and OpenAI embeddings often misclassified relevant content as irrelevant.

Model	PQUAD	Scientific-Specialized	Organizational Report
Lamma-3.1-70B	93.53	89.81	89.06
Qwen2 Instruct - 72B	91.98	78.88	85.27
Gemma 1.1 it	78.00	74.41	75.11
Qwen2-7B	76.30	60.62	82.73
Lamma-3.1-8B	74.89	76.92	84.55
Lamma3 - 8B	72.03	59.15	81.03

Table 6 Baseline Evaluation of Large Language Models.

In the scientific-specialized dataset, MatinaSRoberta retrieved domain-specific jargon more precisely, likely due to its adeptness with specialized terminology, whereas the LaBSE model struggled with specificity due to general training data bias. See **Table 5**.

In the organizational report dataset, containing formal language and cultural references, MatinaSRoberta again surpassed LaBSE and OpenAI embeddings. This dataset underscored challenges posed by formal language, which MatinaSRoberta addressed effectively, unlike LaBSE, which missed cultural nuances. MatinaSRoberta's consistent high performance illustrates the need for specialized embeddings tailored to Persian's complexities, especially in domain-specific contexts.

6.2 Evaluation of Large Language Models in RAG

6.2.1 Baseline Evaluation

The baseline evaluation considered the LLMs' generative abilities, detailing their performance in Table 6. Larger models like LLaMA-3.1 (70B) and Qwen2 Instruct (72B) showed superior capability across all datasets. In PQuad, they managed diverse topics with high precision, unlike smaller models such as Qwen2 (7B) and Gemma 1.1, which struggled with domain coverage. See Table 6 in the Appendix.

In conclusion, larger LLMs consistently outperform smaller ones in all datasets, highlighting the significance of model size and complexity for precise tasks in domain-specific and formal contexts.

6.2.2 Temperature Testing

Temperature testing on LLaMA-3.1 8B evaluated randomness in text generation (Table 7). A temperature of 0.25 gave the best balance between accuracy and diversity. Higher temperatures increased randomness, reducing precision, especially in formal datasets like organizational reports. See **Table 7**.

Temperature	PQUAD	Scientific-Specialized	Organizational Report
0	73.25	80	83.87
0.25	72.83	82.25	84.87
0.5	72.53	76.92	84.55
0.75	72.10	71.66	81.91

Table 7 Impact of Temperature on Model Performance.

6.2.3 Chunk Size Testing

The chunk size test (Table 8) revealed that smaller chunks (512 tokens) provided better precision, especially in formal datasets. Smaller chunks allowed more precise retrieval, while larger chunks reduced precision, particularly in specialized datasets requiring granular information. See **Table 8**.

Chunk Size	PQUAD	Scientific-Specialized	Organizational Report
512	72.13	75.47	92.62
1024	72.41	78.18	84.31
2048	70.37	65.82	84.38

Table 8 Impact of Chunk Size on Model Performance.

With/Without Summary	PQUAD	Scientific- Specialized	Organizational Report
With Summary	77.47	65.11	87.31
Without Summary	71.87	64.47	84.07

Table 9: Impact of Document Summary Indexing on Model Performance.

Model	Context Precision	Faithfulness	Answer Relevancy	Context Recall
Lamma-3.1-70B	0.7750	0.6357	0.7455	0.9205
Qwen2-Instruct-72B	0.7750	0.6125	0.7026	0.9205
Gemma 1.1 it	0.7750	0.5410	0.3779	0.9205
Gemma 2 it	0.7750	0.5524	0.4234	0.9205
Qwen2-7B	0.7750	0.5500	0.6535	0.9205
Lamma-3.1-8B	0.7750	0.6412	0.7233	0.9205
Lamma3-8B	0.7750	0.6313	0.6938	0.9205

Table 10 PQUAD RAGAS.

Model	Context Precision	Faithfulness	Answer Relevancy	Context Recall
Lamma-3.1-70B	0.8138	0.7322	0.6317	0.8314
Qwen2-Instruct-72B	0.8138	0.6928	0.5728	0.8314
Gemma 1.1 it	0.8138	0.6035	0.4656	0.8314
Gemma 2 it	0.8138	0.6142	0.4452	0.8314
Qwen2-7B	0.8138	0.6101	0.4835	0.8314
Lamma-3.1-8B	0.8138	0.6412	0.7233	0.8314
Lamma3-8B	0.8138	0.6313	0.5151	0.8314

Table 11 Scientific-Specialized RAGAS.

Model	Context Precision	Faithfulness	Answer Relevancy	Context Recall
Lamma-3.1-70B	0.5355	0.7388	0.7507	0.8014
Qwen2-Instruct-72B	0.5355	0.7081	0.7214	0.8014
Gemma 1.1 it	0.5305	0.6173	0.3819	0.8014
Gemma 2 it	0.5305	0.6371	0.2919	0.8014
Qwen2-7B	0.5355	0.6942	0.6166	0.8014
Lamma-3.1-8B	0.5305	0.6926	0.7103	0.8014
Lamma3-8B	0.5355	0.5548	0.6085	0.8014

Table 12 Scientific-Specialized RAGAS.

6.2.4 Document Summary Index Testing

Document summary indexing considerably enhanced retrieval accuracy in general and formal datasets (Table 9). This approach improved retrieval efficiency and accuracy, especially for complex queries necessitating information from multiple sections. See **Table 9**.

6.3 RAGAS Results

Since **Context Precision** and **Context Recall** purely evaluate the **retriever** component, their values remain largely consistent across all models using the same retrieval setup. Therefore, our analysis focuses on the generation-related metrics.

The RAGAS framework evaluated retrieval and generation across all datasets. LLaMA-3.1 (70B) consistently showed superior performance in answer relevance across all datasets. In PQuad (Table 10), it recorded the highest answer relevancy (0.7455). Smaller models like Gemma 1.1 struggled, particularly with answer coherence.

In scientific-specialized datasets (Table 11), LLaMA-3.1 (70B) led with the highest faithfulness (0.7322). Though LLaMA-3.1 (8B) produced factually inconsistent responses, its overall semantic alignment remained strong. Gemma models had difficulties with specialized language.

For organizational reports (Table 12), LLaMA-3.1 (70B) again showed strong answer relevancy (0.7507) and faithfulness (0.7388). Gemma models underperformed, notably in relevance.

In summary, LLaMA-3.1 (70B) consistently outperformed in all datasets, indicating robust retrieval capability and high accuracy. The Qwen2 models followed but struggled with specialized content. Gemma models excelled in retrieval but faltered in generating relevant, factual responses in complex contexts, stressing the need for further optimization for smaller models in formal and domain-specific tasks in Persian. See **Tables 10, 11, and 12**.

7. Conclusion

This study addressed the significant challenges associated with extending Retrieval-Augmented Generation (RAG) frameworks to low-resource languages, with a specific focus on Persian. Recognizing the linguistic complexities and the scarcity of annotated datasets, this research aimed to develop Persian-specific language models, establish comprehensive benchmarks, and identify best practices for optimizing RAG systems. These contributions hold critical value in advancing NLP for underrepresented languages and expanding the applicability of retrieval-enhanced frameworks.

The research presented the development of MatinaRoberta and MatinaSRoberta models, designed to capture the nuances of Persian's rich morphology and syntax. These models demonstrated superior performance in retrieval tasks compared to existing state-of-the-art embeddings. The introduction of a systematic benchmark provided a robust means of evaluating RAG systems across general, technical, and formal domains, further highlighting the models' ability to bridge gaps in existing NLP resources. Additionally, the results underscored the importance of tailoring model architectures and retrieval-generation pipelines to meet the demands of low-resource languages, offering insights transferable to other underrepresented linguistic contexts.

The implications of these findings are manifold. Theoretically, they contribute to the growing body of knowledge on enhancing RAG systems through language-specific adaptations. Practically, they lay the groundwork for improving domain-specific applications, such as search engines, legal document analysis, and educational content retrieval in Persian. Moreover, the insights derived from model evaluations and configurations such as temperature tuning, chunk size optimization, and document summary indexing provide actionable guidelines for practitioners and policymakers seeking to enhance RAG systems in resource-

constrained environments.

In conclusion, this research marks a significant step forward in addressing the linguistic and computational challenges of applying RAG systems to low-resource languages like Persian. By combining innovative model development with rigorous evaluation frameworks, the study enhances the understanding of RAG in such contexts and provides a foundation for ongoing innovation and development in the field.

8. Limitations

While the proposed models and benchmarks achieved promising results, several limitations should be acknowledged. The datasets primarily cover standard Persian and therefore may not fully capture dialectal variations or informal language use. Retrieval quality can vary across domains with sparse or unevenly distributed data resources. Additionally, computational constraints limited large-scale experimentation with multilingual or multimodal RAG settings. Addressing these limitations in future work through expanded corpus coverage, bias analysis, and cross-lingual evaluation would further enhance the robustness and generalizability of the proposed approaches.

9. Ethics Statement

This research was conducted in accordance with established ethical standards for computational linguistics. All datasets employed were publicly available or derived from sources explicitly permitting academic use, ensuring full compliance with data licensing, copyright, and privacy requirements. No personally identifiable information or sensitive content was included at any stage of model training or evaluation. The study's focus on Persian as a low-resource language reflects a broader commitment to inclusivity and the equitable advancement of NLP technologies. Throughout development, care was taken to minimize potential biases and to ensure that model outputs do not reinforce social or cultural stereotypes.

10. Bibliographical References

Abbasi, M. A., Ghafouri, A., Firouzmandi, M., Naderi, H., and Minaei Bidgoli, B. (2023). PersianLLaMA: Towards building first Persian large language model. arXiv preprint arXiv:2312.15713.

Bourbour Hosseinbeigi, S., Taherinezhad, F., Faili, H., Baghbani, H., Nadi, F., and Amiri, M. (2025). Matina: A large-scale 73B token Persian text corpus. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Albuquerque, New Mexico, pp. 9143–9157. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.462

Chen, J., Lin, H., Han, X., and Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 17754–17762.

Chowdhary, K. (2020). Natural language processing. In Fundamentals of Artificial Intelligence, pp. 603–649.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020).

Darvishi, K., Shahbodaghkhan, N., Abbasiantaeb, Z., and Momtazi, S. (2023). *PQuAD: A Persian question answering dataset.* Computer Speech & Language, vol. 80, p. 101486.

Dubey, A., et al. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 878–891.

Ghafouri, A., Abbasi, M. A., and Naderi, H. (2023). *AriaBERT: A Pre-trained Persian BERT Model for Natural Language Understanding.* Preprint (Iran University of Science and Technology).

Gao, Y., et al. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Khashabi, D., et al. (2021). Parsinlu: A suite of language understanding challenges for Persian. Transactions of the Association for Computational Linguistics, 9, 1147–1162.

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., and Chen, E. (2025). CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. ACM Transactions on Information Systems (TOIS), 43(2), Article 41, 32 pages. https://doi.org/10.1145/3701228

Mazyar, M. S. and Azaryan, M. M. (2018). General Physical Education. University of Jiroft.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., et al. (2024). Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large Language Models: A Survey. arXiv preprint arXiv:2402.06196.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4996–5001). Florence, Italy.

Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. (2020). DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference

on Knowledge Discovery & Data Mining, pp. 3505–3506.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Le Lan, C., Jerome, S., Tsitsulin, A., Vieillard, N., et al. (2024). Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

SadraeiJavaheri, M., et al. (2024). TookaBERT: A Step Forward for Persian NLU. arXiv preprint arXiv:2407.16382.

Shahul, E., James, J., Anke, L.E., & Schockaert, S. (2023). RAGAs: Automated Evaluation of Retrieval Augmented Generation. Conference of the European Chapter of the Association for Computational Linguistics.

Tonmoy, S., et al. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313.

Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, T., et al. (2024). Searching for best practices in retrieval-augmented generation. In Proceedings of EMNLP 2024.

Yang, A., et al. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

11. Language Resource References

Bourbour Hosseinbeigi, S., Taherinezhad, F., Faili, H., Baghbani, H., Nadi, F., and Amiri, M. (2025). Matina: A large-scale 73B token Persian text corpus. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 9143–9157. Albuquerque, New Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.462

Darvishi, K., Shahbodaghkhan, N., Abbasiantaeb, Z., and Momtazi, S. (2023). PQuAD: A Persian question answering dataset. Computer Speech & Language, 80, 101486.

Khashabi, D., et al. (2021). ParsiNLU: A suite of language understanding challenges for Persian. Transactions of the Association for Computational Linguistics, 9, 1147–1162.

Conneau, A., et al. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116. pid: https://huggingface.co/docs/transformers/model_doc/xlm-roberta

SadraeiJavaheri, M., et al. (2024). TookaBERT: A step forward for Persian NLU. arXiv preprint arXiv:2407.16382.

Mesnard, T., et al. (2024). Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295. pid: https://deepmind.google/models/gemma/

Dubey, A., et al. (2024). Llama 3 herd of models. arXiv preprint arXiv:2407.21783. pid: https://huggingface.co/meta-llama/

Yang, A., et al. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671. pid: https://huggingface.co/collections/Qwen/qwen2-6659360b33528ced941e557f