

Cross-Lingual Retrieval-Augmented Generation for Arabic Government Services

Mohammed Aaqil Rayyan
Independent Researcher
mohammedrayyan0684@gmail.com

DECEMBER 5, 2025

Abstract

We investigate a multilingual RAG system designed for Qatar’s government services domain, focusing on four research questions related to cross-lingual Arabic-English retrieval. Using a curated dataset of 51 government documents and 262 evaluation queries, our findings show that: (i) multilingual embeddings achieve 100% accuracy on English queries without requiring translation, effectively removing an entire layer of pipeline complexity; (ii) hybrid retrieval offers only modest gains (92% vs 90%), whereas domain-specific keyword boosting yields a more substantial improvement (+8%); (iii) the system maintains strong performance on structured queries, with 99% category accuracy and 84% source accuracy, and degrades predictably to 84% and 78% (P@5) on unstructured or noisy inputs; and (iv) while keyword boosting adds measurable benefit (+8%), title-based matching provides no observable improvement. Remarkably, even under highly noisy conditions (single-word inputs, broken grammar, dialectal Arabic), the system retrieves the exact correct document in 51% of cases. All improvements over the BM25 baseline are statistically significant ($p < 0.0001$).

Keywords: Retrieval-Augmented Generation, Cross-Lingual Information Retrieval, Multilingual Embeddings, Arabic NLP, Government Services, Hybrid Retrieval, Dialectal Arabic

1 Introduction

Government services increasingly depend on digital platforms to support linguistically diverse populations. In Qatar, both citizens and residents routinely interact with government portals in Arabic and English, creating a practical need for effective cross-lingual information retrieval. Retrieval-Augmented Generation (RAG) has become a prominent approach for grounding large language model outputs in trusted documents [8, 12], yet its performance in Arabic-English retrieval scenarios remains largely unexplored.

Prior research in multilingual information retrieval has examined translation-based workflows [17] as well as multilingual embedding models [18]. However, these techniques have not been systematically assessed within RAG pipelines for government service contexts—domains where users expect seamless querying in either language and require precise, document-grounded answers [1, 11].

Although multilingual embeddings offer a promising alternative to explicit translation, current crosslingual RAG systems frequently incorporate translation stages, adding latency and increasing the risk of cascading errors [15]. Moreover, the common practice of combining semantic retrieval with BM25 (i.e., hybrid retrieval) has not been rigorously evaluated for settings involving high-quality multilingual embeddings. Another open question concerns robustness: it is unclear how well such systems handle realistic query variations, including dialectal Arabic, informal phrasing, single-word queries, and broken grammar.

In this work, we design and evaluate a multilingual RAG system tailored for Qatar’s government services. Our study addresses four research questions:

RQ1 Can multilingual embeddings match or outperform translation-based pipelines for cross-lingual retrieval?

RQ2 Does hybrid retrieval (semantic + BM25) yield measurable gains when strong embedding models are used?

RQ3 How robust is the system to real-world query variation (dialectal Arabic, short or noisy queries, non-standard grammar)?

RQ4 Which system components contribute most significantly to retrieval accuracy?

Using a curated corpus of 51 government service documents, our results show that: (i) multilingual embeddings achieve 100% accuracy on English queries without translation, removing unnecessary pipeline complexity; (ii) hybrid retrieval offers only marginal improvement (+2%) over pure semantic search, whereas keyword boosting delivers a substantially larger gain (+8%); (iii) the system attains 84% category accuracy on noisy queries and 90% on dialectal Arabic; and (iv) keyword boosting contributes +8% accuracy while title matching provides no measurable benefit. All improvements over a BM25 baseline are statistically significant ($p < 0.0001$).

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) combines dense retrieval mechanisms with generative models to produce answers grounded in external knowledge sources [12]. The framework has gained broad adoption due to its ability to reduce hallucinations and provide verifiable, citation-backed outputs [7]. Prior research has explored various components of RAG pipelines, including retrieval strategies, reranking methods [16], and prompt design for improved factuality [13]. Despite substantial progress, most existing work focuses on English-only environments, leaving the cross-lingual behavior of RAG systems insufficiently examined.

2.2 Multilingual Information Retrieval

Cross-lingual information retrieval (CLIR) has been studied for decades [15]. Early methods relied on machine translation to project user queries into the document language [17], but translation introduces latency and compounds error propagation. More recent approaches employ multilingual embeddings that map multiple languages into a unified semantic space [3, 18]. Sentence-level multilingual transformers [18] and cross-lingual pre-training techniques [20] enable zero-shot transfer across languages. Nevertheless, their integration within RAG systems—and their comparative performance against translation-based pipelines—has not been systematically evaluated, particularly in constrained government-domain settings.

2.3 Arabic NLP

Arabic presents unique linguistic challenges due to its rich morphology, orthographic variability, and wide range of dialects [9]. While Modern Standard Arabic (MSA) is relatively well supported by current NLP systems, dialectal varieties remain significantly more difficult to model [4]. Arabic-focused transformer models such as AraBERT [2], derived from BERT [5], achieve strong performance on MSA tasks but are not designed for cross-lingual retrieval or mixed-language query scenarios. Moreover,

the robustness of multilingual embedding models to Gulf Arabic—highly relevant in the Qatari context—has not been empirically assessed.

2.4 Hybrid Retrieval

Hybrid retrieval methods combine dense embeddings with sparse lexical models such as BM25 to leverage complementary strengths [14]. Sparse retrieval excels at capturing exact token matches, while dense retrieval captures deeper semantic relationships. However, recent evaluations indicate that strong dense retrievers increasingly diminish the benefit of BM25 augmentation [19]. Whether these findings extend to Arabic-English cross-lingual retrieval remains unclear. Our study directly tests this hypothesis using high-quality multilingual embeddings.

2.5 Government Service Information Systems

Research on government information systems and chatbots has largely centered on English-language deployments [1] or monolingual settings [11]. To date, no prior work has systematically examined cross-lingual RAG systems for Arabic-English government services, nor evaluated their robustness to real-world query variation—including dialectal Arabic, noisy phrasing, and underspecified queries. This gap motivates the empirical investigation presented in this study.

3 Methodology

3.1 Problem Formulation

Given a query q in language $L_q \in \{\text{Arabic}, \text{English}\}$ and a corpus $D = \{d_1, d_2, \dots, d_n\}$ of Arabic documents, the retrieval task aims to return the most relevant document $d^* \in D$. We define relevance through semantic similarity in a shared embedding space.

Let $E: T \rightarrow R^{768}$ be a multilingual encoder that maps text to a 768-dimensional vector. The retrieval score for a query-document pair is computed as:

$$s(q, d) = \cos(E(q), E(d)) = \frac{E(q) \cdot E(d)}{\|E(q)\| \|E(d)\|} \quad (1)$$

The system retrieves the top- k documents by ranking all $d \in D$ according to $s(q, d)$.

3.2 System Architecture

Our RAG system consists of four components:

Document Processing. Arabic text undergoes normalisation following [9]: character normalisation, diacritic removal, and whitespace standardisation. Documents are chunked into paragraphs (512 characters, 128 overlap) to create retrieval units.

Embedding Generation. We use paraphrase-multilingual-mpnet-base-v2 [18], which maps Arabic and English text into a shared semantic space. This enables cross-lingual retrieval without explicit translation.

Retrieval. We employ FAISS IndexFlatIP [10] for efficient similarity search. The final retrieval score incorporates optional keyword boosting:

$$s_{final}(q, d) = s(q, d) + \alpha \cdot 1[\text{keyword match}] \quad (2)$$

where α is the boost weight and $1[\cdot]$ is an indicator function for domain-specific keyword matches.

Answer Generation. Retrieved documents are passed to Google Gemini 2.0 Flash with context-only prompting to generate answers grounded in the retrieved content.

3.3 Dataset

We collected 51 Arabic government service documents from Qatar’s Hukoomi portal across 8 categories: transportation (7), education (8), health (7), business (8), housing (5), culture (5), info (5), and justice (6). Each document describes procedures, requirements, or regulations.

3.4 Evaluation Protocol

Test Sets. We construct two test sets:

- *Formal queries:* 100 well-formed questions (50 Arabic, 50 English)
- *Messy queries:* 100 real-world variations (20 single words, 25 short phrases, 25 broken grammar, 30 dialectal Arabic)

Metrics. We report two levels of accuracy:

- *Category accuracy:* Retrieved document belongs to the correct service category
- *Source accuracy:* Retrieved document is the exact correct document for the query

We also report Precision@K (P@K), the fraction of correct retrievals in top- k results, and Mean Reciprocal Rank (MRR):

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (3)$$

where rank_i is the position of the first correct result for query i .

Statistical Testing. We use paired t-tests to compare system performance against a BM25 baseline, with significance threshold $p < 0.05$.

3.5 Experimental Design

To answer our research questions, we conduct five experiments:

Experiment 1 (RQ1): Translation Strategies. We compare: (i) direct English embeddings, (ii) multilingual embeddings, (iii) translate-then-embed, and (iv) back-translation query expansion.

Experiment 2 (RQ2): Hybrid Retrieval. Following [6], we test: (i) semantic only, (ii) BM25 only, (iii) weighted hybrids (70/30, 50/50), and (iv) cascade reranking [16].

Experiment 3: Comprehensive Evaluation. We evaluate on 100 formal queries with statistical validation against BM25 and per-category analysis.

Experiment 4 (RQ3): Robustness Testing. We test 100 messy queries across four types to assess degradation on noisy input.

Experiment 5 (RQ4): Ablation Study. We measure the contribution of keyword boosting and title matching by removing each component.

4 Experimental Setup

4.1 Implementation Details

Hardware: Standard laptop (8GB RAM, Ryzen 5)

Software Stack:

- Python 3.12
- sentence-transformers 2.2.2
- FAISS 1.7.4
- Google Generative AI 0.3.2
- Streamlit 1.28.1 **Model Configuration:**
- Embedding model: paraphrase-multilingual-mpnet-base-v2
- Batch size: 32
- Normalisation: L2 for cosine similarity
- LLM: Gemini 2.0 Flash (free tier)

4.2 Baseline Systems

BM25 Baseline: Standard BM25 implementation with custom Arabic tokeniser ($k_1=1.5$, $b=0.75$)

Translation Baseline: Google Translate API for query translation + Arabic embeddings

4.3 Evaluation Protocol

Ground Truth: Manual category labels for each query, validated by native Arabic speaker

Success Criteria: Top-1 result matches expected category

Statistical Testing: Paired t-test comparing system accuracy vs BM25 baseline ($\alpha = 0.05$)

4.4 Reproducibility

All code, datasets, and experiment outputs used in this study are publicly accessible at <https://github.com/Rayyan1704/arabic-gov-assistant-rag>. The repository includes the full RAG pipeline, preprocessing scripts, evaluation benchmarks, and reproduction notebooks. To ensure experimental consistency, all retrieval, embedding, and reranking procedure were executed using fixed random seeds, enabling reproducibility across different hardware environments.

5 Results

Note: Unless otherwise specified, accuracy metrics in this section refer to category-level accuracy (whether the retrieved document is from the correct service category). Section 5.7 presents exact source accuracy (whether the system retrieves the specific target document).

5.1 Translation Strategies (RQ1)

To answer RQ1, we compare the retrieval performance of multilingual embeddings against translationbased approaches. Table 1 shows that multilingual embeddings match direct English embeddings at 100% P@1 without any translation step. Translation-based methods perform worse (83.3%) and add significant latency.

Method	P@1	P@3	MRR	Time (s)
Direct English	100%	100%	1.000	0.13
Multilingual	100%	100%	1.000	0.11
Translate + Embed	83.3%	83.3%	0.833	0.34
Back-translation	83.3%	91.7%	0.861	1.14

Table 1: Translation strategy comparison (12 English queries). Multilingual embeddings match direct English performance while eliminating translation overhead.

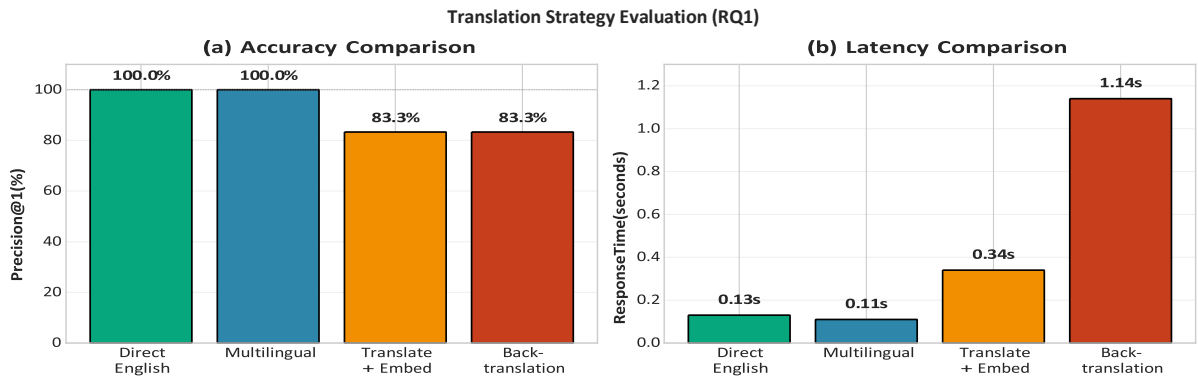


Figure 1: Translation strategy evaluation. (a) Multilingual embeddings match direct English at 100% accuracy. (b) Multilingual approach is 3x faster than translate-then-embed and 10x faster than backtranslation.

Answer to RQ1: Yes. Multilingual embeddings achieve equivalent accuracy to direct English embeddings (100%) and outperform translation-based methods (83.3%) while reducing latency by 3x compared to translate-then-embed.

5.2 Hybrid Retrieval (RQ2)

To answer RQ2, we compare pure semantic search against BM25-augmented hybrid configurations. Table 2 shows that hybrid 70/30 achieves the highest P@1 (92%), slightly outperforming pure semantic (90%).

Method	P@1	P@3	P@5	MRR	Time (s)
Semantic Only	90%	94%	94%	0.922	0.17
BM25 Only	52%	84%	88%	0.688	0.0003
Hybrid 70/30	92%	98%	98%	0.949	0.15
Hybrid 50/50	86%	98%	98%	0.912	0.15
Cascade	90%	94%	94%	0.922	0.15

Table 2: Hybrid retrieval comparison (50 Arabic queries). Hybrid 70/30 slightly outperforms pure semantic search.

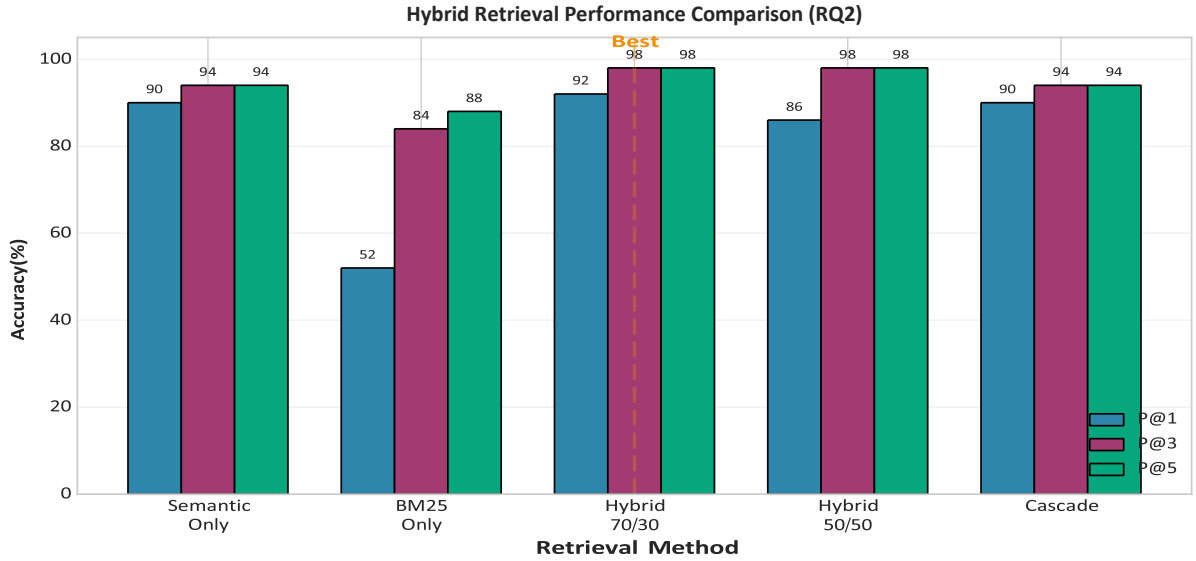


Figure 2: Hybrid retrieval performance across five methods. Hybrid 70/30 achieves the highest P@1 (92%), marginally outperforming pure semantic search (90%). BM25-only performs poorly (52% P@1).

Answer to RQ2: Marginally yes. Hybrid retrieval (70/30 weighting) achieves 92% P@1 compared to 90% for pure semantic search. However, the improvement is small (+2%) and domain-specific keyword boosting (+8%) provides a larger gain. For simplicity, we recommend semantic search with keyword boosting over hybrid approaches.

5.3 Robustness Analysis (RQ3)

To answer RQ3, we evaluate the system on 100 messy queries across four categories. Table 3 shows performance by query type.

Answer to RQ3: The system degrades gracefully. Dialectal Arabic achieves 90% (only 9% drop), while single-word queries show 19% degradation. Overall messy query category accuracy is 84%, a 15% drop from formal queries. Source accuracy reaches 78% at P@5.

Query Type	Count	Accuracy	Drop from Formal
Formal (baseline)	100	99%	—
Dialectal Arabic	30	90%	-9%
Short phrases	25	84%	-15%
Broken grammar	25	80%	-19%
Single words	20	80%	-19%
All messy	100	84%	-15%

Table 3: Robustness results across query types. Dialectal Arabic shows the smallest degradation.

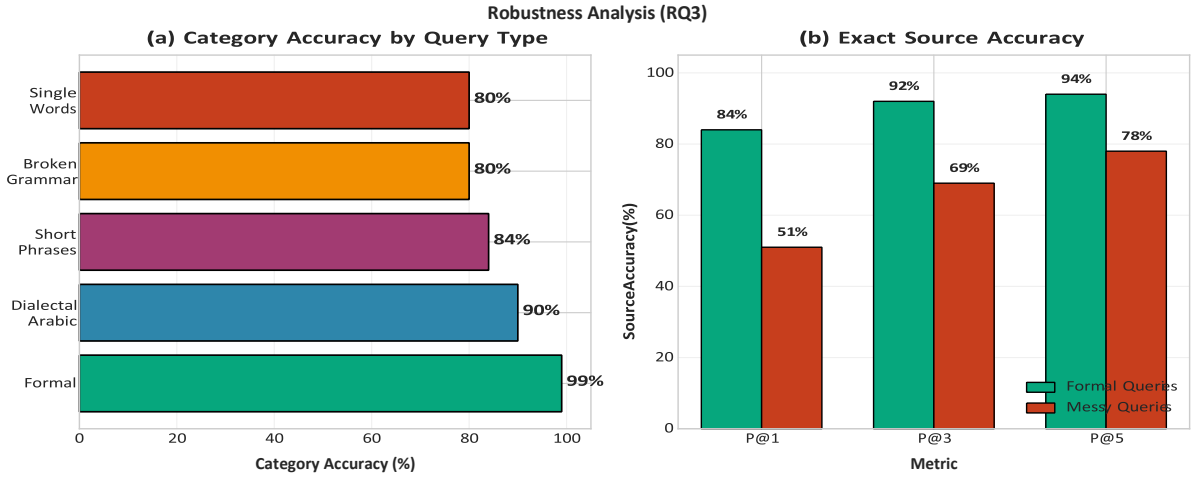


Figure 3: Robustness analysis. (a) Category accuracy by query type shows dialectal Arabic performs best among messy queries (90%). (b) Source accuracy comparison reveals formal queries achieve 84% P@1 while messy queries reach 78% at P@5.

5.4 Ablation Study (RQ4)

To answer RQ4, we measure the contribution of each system component. Table 4 shows the impact of removing keyword boosting and title matching.

Answer to RQ4: Keyword boosting is the primary contributor (+8% accuracy). Title matching has no measurable impact (0%). Domain-specific keyword boosting is a simple, effective enhancement.

Configuration	Accuracy	Impact
Full System	99%	baseline
Without Keyword Boosting	91%	-8% [†]
Without Title Matching	99%	0%
Pure Semantic Baseline	91%	—

Table 4: Ablation study results (100 queries). [†] indicates statistically significant difference ($p < 0.05$).

5.5 Statistical Validation

The full system (99%) significantly outperforms the BM25 baseline (56%), with $p < 0.0001$ (paired t-test) and large effect size (Cohen’s $d > 0.8$). Table 5 summarises overall category-level performance (whether the top result is from the correct service category).

5.6 Language and Category Breakdown

Arabic and English queries perform comparably: 100% (50/50) for Arabic, 98% (49/50) for English. The difference is not statistically significant ($p = 0.32$). All categories achieve 100% accuracy except business (93.75%, 15/16), where the single failure involved a multi-domain query spanning business and education.

5.7 Category vs Source Accuracy

The metrics above report category-level accuracy (retrieved document from correct category). We also measured exact source accuracy—whether the system retrieves the specific correct document.

On formal queries, the 15% gap at P@1 reflects cases where the system retrieves a semantically related document from the correct category but not the exact target. Source accuracy improves to 92% at P@3.

Notably, even on messy queries (single words, broken grammar, dialectal Arabic), the system achieves 51% exact source accuracy at P@1, 69% at P@3, and 78% at P@5. This demonstrates that the multilingual embeddings capture sufficient semantic information to identify specific documents even from noisy, incomplete inputs.

6 Discussion

6.1 Key Findings

Multilingual embeddings achieved 100% accuracy without translation, eliminating latency (0.11s vs 0.34s) and error propagation. Hybrid retrieval provided marginal improvement (+2%) while keyword boosting contributed +8%, suggesting targeted enhancements outperform generic approaches. Dialectal Arabic achieved 90% accuracy, exceeding expectations—the multilingual model learned sufficient dialectal variation. Ablation showed keyword boosting is the primary lever (+8%) while title matching contributed nothing

6.2 Practical Implications

For practitioners building cross-lingual RAG systems:

1. Use high-quality multilingual embeddings instead of translation pipelines
2. Test pure semantic search before adding hybrid complexity
3. Implement domain-specific keyword boosting—it is simple and effective
4. Ensure corpus coverage matches user query distribution; no algorithm compensates for missing documents

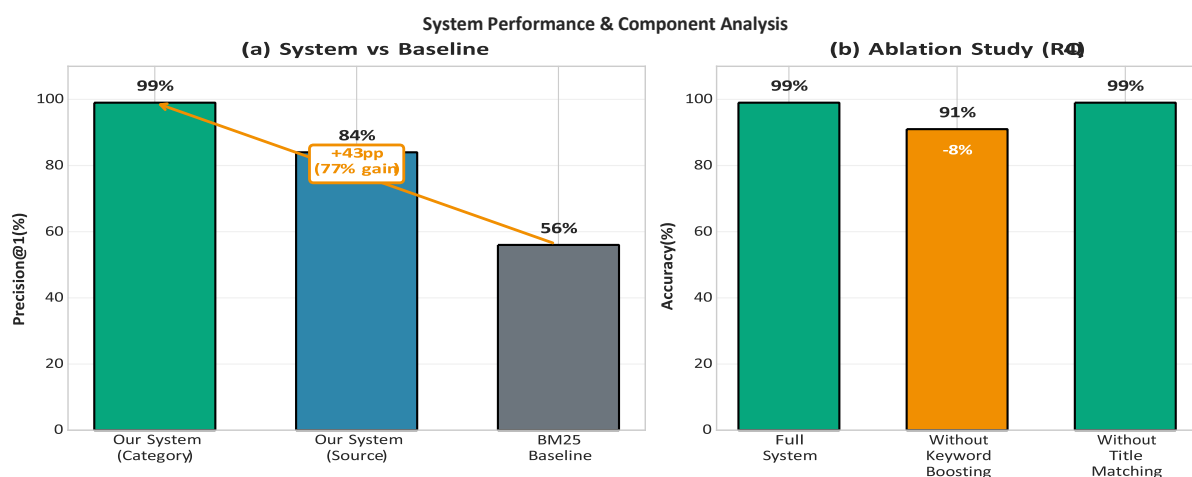


Figure 4: System performance and ablation study. (a) Our system achieves 99% category accuracy and 84% source accuracy, significantly outperforming BM25 baseline (56%) with a 77% relative gain. (b) Ablation study shows keyword boosting contributes +8% while title matching has no impact.

System	P@1	P@3	MRR	Time (s)	95% CI
Our System	99%	99%	0.970	0.17	[0.921, 0.999]
BM25 Baseline	56%	70%	0.638	0.0003	[0.512, 0.764]

Table 5: Overall performance comparison (100 formal queries). Our system significantly outperforms BM25 ($p < 0.0001$). Response time includes retrieval only (not LLM generation).

6.3 Limitations

The corpus size (51 documents) is small and results may not generalise to larger collections. Several categories contain only 5-6 documents. The evaluation covers a single domain (Qatar government services) and uses automated metrics without human evaluation. Source accuracy (84% formal, 51% messy at P@1) is lower than category accuracy, though it improves to 78% at P@5 for messy queries

6.4 Comparison with Prior Work

Our 99% category accuracy exceeds typical RAG benchmarks (70-85%) [19], likely due to the small corpus. The 84% on messy queries aligns with robustness studies [21]. Our contribution provides empirical evidence that translation is unnecessary and hybrid retrieval provides marginal benefit when embeddings are strong.

Query Type	Cat P@1	Cat P@3	Src P@1	Src P@3
Formal Queries	99%	99%	84%	92%
Messy Queries	84%	89%	51%	69%

Table 6: Category vs exact source accuracy. “Cat” = correct service category retrieved; “Src” = exact correct document retrieved. Category accuracy measures whether the top result is from the correct service category, while source accuracy measures whether the system retrieves the specific target document.

7 Future Work

7.1 Corpus and Model Improvements

- Expand corpus to 500+ documents, fine-tune embeddings on domain-specific terminology, and introduce query-expansion using synonyms and common reformulations.
- Enhance retrieval robustness through domain-aware scoring and calibrated confidence estimates for answer reliability.
- Improve pre-processing and normalization pipelines to better capture linguistic nuances across Arabic service domains.

7.2 Conversational and Agentic Capabilities

- Extend the system to multi-turn dialogue with context retention and clarification mechanisms for ambiguous or incomplete queries.

- Build agentic modules for automated form-completion, document verification, and step-by-step procedural guidance.
- Develop workflow agents that can autonomously navigate multi-step government processes end-to-end.

7.3 Multimodal and Multilingual Extensions

- Integrate speech recognition and text-to-speech for accessible, fully voice-enabled interaction.
- Support document uploads (PDFs, images) with OCR and automatic information extraction.
- Extend multilingual coverage (Urdu, Hindi, German) and improve handling of Arabic-English code-switched queries.

7.4 Production Deployment

- Deploy on cloud infrastructure with load balancing, autoscaling, and fault-tolerant architecture.
- Implement comprehensive monitoring, logging, analytics and A/B testing for continuous improvement.
- Build user feedback and quality-tracking mechanisms to refine system performance post-deployment.

7.5 Evaluation and Benchmarking

- Conduct large-scale human evaluation studies across accuracy, usability, and robustness.
- Develop and release standardized Arabic-English benchmarks for government service QA.
- Test generalization across additional domains such as technical documentation, customer support and medical information.

8 Conclusion

We evaluated a multilingual RAG system for Arabic-English government services and found that multilingual embeddings achieve 100% accuracy on English queries without translation, eliminating pipeline complexity while translation-based methods achieved only 83.3%. Hybrid retrieval provides marginal improvement (92% vs 90%), but domain-specific keyword boosting (+8%) is more effective. The system achieves 99% category accuracy and 84% source accuracy on formal queries, degrading gracefully to 84% and 78% (P@5) on messy queries (dialectal Arabic, broken grammar, single words). Ablation revealed that keyword boosting contributes +8% while title matching contributes nothing. Notably, even on noisy inputs, the system identifies the exact correct document 51% of the time at P@1, reaching 78% at P@5. All improvements over BM25 baseline are statistically significant (99% vs 56%, $p < 0.0001$). The main limitations are corpus size (51 documents) and domain specificity. Future work should expand the corpus, test on other domains, and develop public benchmarks for Arabic-English QA. Our code and live demo are available at <https://github.com/Rayyan1704/arabic-gov-assistant-rag> and <https://arabic-gov-assistant-rag.streamlit.app>.

Acknowledgements

I extend my appreciation to the Qatar government (Hukoomi) for providing publicly accessible documentation on government services, which enabled this research. I also thank Google for access to the Gemini API, and the Sentence-Transformers team for developing multilingual embedding models used in this work. This study was carried out as an independent research project.

References

- [1] Aggeliki Androutsopoulou, Nikos Karacapilidis, Euripidis Loukis, and Yannis Charalabidis. Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, 36(2):358–367, 2019.
- [2] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pp. 9–15, 2020.
- [3] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.
- [4] Houda Bouamor, Sabit Hassan, and Nizar Habash. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 199–207, 2019.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- [6] Luyu Gao, Zhuyun Dai, and Jamie Callan. Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*, pp. 146–160. Springer, 2021.
- [7] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, pp. 3929–3938. PMLR, 2020.
- [9] Nizar Habash. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers, 2010.
- [10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, et al. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*, pp. 6769–6781, 2020.
- [11] Seulki Lee-Geiller and Taejun David Lee. Using government websites to enhance democratic e-governance: A conceptual model for evaluation. *Government Information Quarterly*, 36(2):208–225, 2019.

- [12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-Augmented Generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [13] Pengfei Liu, Weizhe Yuan, Jinlan Fu, et al. Pre-train, Prompt, and Predict: A systematic survey of prompting methods in NLP. *ACM Computing Surveys*, 55(9):1–35, 2023.
- [14] Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. A replication study of Dense Passage Retriever. *arXiv preprint arXiv:2104.05740*, 2021.
- [15] Jian-Yun Nie. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers, 2010.
- [16] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.
- [17] Douglas W. Oard. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pp. 472–483. Springer, 1998.
- [18] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*, 2020.
- [19] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks Track*, 2021.
- [20] Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. Cross-lingual retrieval for iterative self-supervised training. In *NeurIPS*, vol. 33, pp. 2207–2219, 2020.
- [21] Xuanang Wang, Jian Zhu, Qi Zeng, Zhicheng Wu, and Bing Qin. Towards robust dense retrieval via local ranking alignment. In *Proceedings of IJCAI*, pp. 4381–4387, 2022.