Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 533-541.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Robust Multilingual RAG under Query Perturbations: An English-Persian Benchmark

Niloofar Ranjbar* and Hamed Baghbani

Jam Technical Faculty, Persian Gulf University, Jam, Iran.

Article Info

Article History:
Received 14 March 2026
Revised 08 July 2026
Accepted 15 July 2026

DOI:10.22044/jadm.2026.17608.2908

Keywords:
Retrieval-augmented Generation; Multilingual Information Retrieval, Robustness Evaluation, English-Persian Retrieval, Hybrid Retrieval, Benchmark.

**Corresponding author: nranjbar@pgu.ac.ir (N. Ranjbar).*

Abstract

Retrieval-augmented generation (RAG) is commonly evaluated on clean inputs that underrepresent realistic multilingual variation. We present an English-Persian movie-domain robustness benchmark built from a corpus of 31,564 records, 120 clean queries, and 720 aligned perturbations. The benchmark covers six deterministic query types and 14 operational perturbation labels grouped into four families. We compare BM25, multilingual dense retrieval, character n-gram TF-IDF, and hybrid retrieval, and evaluate top-1 deterministic answer extraction against a field-specific top-5 RAG system using Qwen2-7B-Instruct. Hybrid retrieval achieves 81.50 MRR@10 on clean queries and 67.76 under perturbation; field-specific RAG reaches 84.17% and 72.08% accuracy, respectively. Clustered paired-bootstrap 95% confidence intervals exclude zero for all principal system differences. English-title noise is the most damaging family, whereas query-form and punctuation variation are comparatively well tolerated. A 43-case consistency audit verifies the implementation of the rule-based failure categories, and full-output analysis shows that retrieval-coverage errors dominate the difficult English-title family. These results support component-level evaluation of multilingual RAG robustness.

1. Introduction

Retrieval-augmented generation (RAG) grounds large language models in external evidence by coupling retrieval with generation [1]. This paradigm builds on retrieval-centered work such as Dense Passage Retrieval and KILT [2,3] and is now evaluated across question answering and diverse domain-specific settings [4-6]. Strong clean-input performance, however, does not guarantee robustness when queries contain spelling errors, spacing variation, punctuation loss, script confusion, or cross-lingual ambiguity [7-9].

The problem is especially important in mixed English-Persian retrieval. Users may alternate between scripts and languages and may introduce Arabic-Persian character substitutions, half-space variation, transliteration, or keyboard-induced corruption. These changes can degrade both sparse and dense retrieval, after which errors propagate into extraction or generation [7,8]. MIRACL further demonstrates substantial variation across languages and scripts [10], while query-level RAG studies motivate separating retrieval and downstream failures [11].

We study this problem using a Persian-centered movie corpus of 31,564 records and a manually curated benchmark of 120 clean queries. The same 20 underlying bilingual movies are used across six query types, producing 20 examples per type; each clean query has six meaning-preserving variants, yielding 720 perturbed queries. Each query is linked to a query-specific gold record using normalized title matching followed by available year and director constraints. The perturbations comprise 14 operational labels grouped into four families: Persian script/spacing, Persian character noise, English-title noise, and query-form/punctuation variation.

The movie domain is a useful controlled test bed because it naturally contains bilingual and alternative titles, named entities, years, directors, punctuation, and mixed-script forms. These properties create realistic lexical anchors and ambiguity, while structured metadata supports unique record linking and deterministic answer scoring.

We compare BM25, multilingual dense retrieval, character n-gram TF-IDF, and a min-max normalized BM25-dense hybrid. On top of the hybrid rankings, we compare a transparent top-1 field-extraction reference with a field-specific Qwen2-7B-Instruct system that receives the top five records. Because the two answering systems receive different evidence depths, their comparison reflects both multi-record context access and model-based evidence selection and generation; it does not isolate generation alone.

We are not aware of an English-Persian benchmark that jointly combines aligned clean and perturbed queries, controlled record retrieval, deterministic and generative answering, perturbation-family analysis, and stage-level failure attribution. The contributions are: (1) a paired English-Persian robustness benchmark; (2) a formal perturbation taxonomy; (3) sparse, dense, character-level, and hybrid retrieval comparisons; (4) hybrid-weight sensitivity and clustered confidence intervals; and (5) rule-based failure attribution with a consistency audit and full-output failure distribution.

2. Related Work

2.1. RAG Evaluation and Query Robustness

RAG evaluation increasingly separates answer correctness from retrieval relevance, evidence use, faithfulness, and refusal. RAGAs, ARES, and RAGBench provide complementary, component-aware evaluation approaches [4-6], while GaRAGe and Trust-Align emphasize grounding and refusal when evidence is insufficient [12,13]. Typo-robust dense retrieval methods show that small character changes can substantially reduce effectiveness [7,8], and BEIR demonstrates uneven retriever generalization across datasets [9]. Perçin et al. directly examine query perturbations across RAG components and motivate stage-level diagnosis [11].

2.2. Sparse, Dense, and Hybrid Retrieval

Word-level sparse retrieval remains effective when titles, years, and names provide exact anchors [9]. Dense bi-encoders improve semantic matching [2], whereas ColBERT and SPLADE preserve richer token-level or learned lexical interactions [14,15]. We include character 3-5-gram TF-IDF as a lightweight, perturbation-oriented baseline because overlapping character sequences can survive some deletions, repetitions, transpositions, and spacing changes. It is not intended as a substitute for late-interaction or learned sparse retrievers.

2.3. Multilingual and Persian RAG

Multilingual retrieval introduces challenges such as script normalization, transliteration, mixed-language queries, and language-specific tokenization. MIRACL documents large cross-language differences [10], MEMERAG studies multilingual RAG meta-evaluation [16], and Li et al. report uneven cross-lingual retrieval and generation [17]. PersianRAG analyzes Persian preprocessing, embeddings, retrieval, prompt construction, language modeling, and evaluation [18]. Bourbour Hosseinbeigi et al. develop Persian-specific models and benchmark retrieval-generation settings across several domains [19]. These works focus on system construction and optimization; our study fixes a controlled corpus and answer space and analyzes paired perturbations and observable pipeline failure stages.

3. Methodology

3.1. Task and Pipeline

Given a natural-language query, the system retrieves a movie record and returns a deterministic field or a constrained multi-field answer. The pipeline consists of corpus preparation, paired benchmark construction, retrieval, and answering. Each benchmark row is linked to a unique gold record and a deterministically scorable answer.

3.2. Corpus Construction

The corpus was assembled from three sources: a Persian Wikipedia/Wikidata-derived JSONL collection; the Iranian Movies dataset [20]; and The Movies Dataset [21]. The Persian source served as the principal backbone, the English metadata source provided bilingual enrichment, and the Iranian movie dataset contributed additional Persian records and metadata.

Records were mapped to a unified schema containing movie identifiers, Persian and English titles, language, year, genres, directors, Persian and English plots, URLs, and optional actors and rating metadata. Persian records were aligned to English metadata with normalized multilingual title embeddings, a +/-1-year candidate window, and a cosine-similarity threshold of 0.82. Unmatched Persian records were retained; unmatched English-only records were excluded from the final

experimental corpus. Cleaning applied Unicode NFKC normalization, Arabic-to-Persian character mapping, diacritic and invisible-character removal, Persian/Arabic digit conversion, and whitespace normalization. Deduplication used normalized title-year keys and retained the metadata-richest record. The retrieval representation concatenated titles, year, genres, directors, actors, ratings, and available summaries. The final corpus contains 31,564 records. Treating null, empty, and literal missing-value placeholders as missing, all records have a usable Persian title, 10,836 have an English title, 31,109 contain Persian plot text, 9,254 contain English plot text, 10,273 include genre metadata, and 29,191 include director information.

Please be sure your sentences are complete and that there is continuity within your paragraphs. Check the numbering of your graphics and make sure that all proper references are included.

3.3. Benchmark Design

The 120 clean queries are generated from 20 underlying bilingual movies. Each movie appears once in each of the six query types, so type-level comparisons use a matched set of movies rather than different item pools. The records were selected to provide bilingual titles, complete year and director metadata, and opportunities for the applicable perturbation operators. Table 1 specifies the input and answer structure.

Table 1. Benchmark query types (FA: Persian; EN: English).

Query type	Input	Required answer	n
Year	FA title	Release year	20
Director	FA title	Director	20
EN-to-FA title	EN title	FA title	20
Year-constrained	EN title + year	FA title	20
Year + director	FA title	Year + director	20
Cross-lingual fields	EN title	Year + director	20

3.3.1. Gold Linking

Rows are linked by exact matching of normalized Persian or English titles, followed by year and director disambiguation. If multiple records remain, a deterministic metadata-completeness score selects the most comprehensive record. Evaluation retains the resulting query-specific exact gold-record identifier; an alternate corpus record for the same underlying movie is not considered equivalent when its record-level metadata differs. All 120 clean and 720 perturbed queries were linked.

3.3.2. Formal Perturbation Model

For each clean instance (q_i, m_i, a_i, τ_i), the perturbation operator T_l generates a surface-form variant while preserving the original information need and maintaining the same gold movie, gold answer, and query type:

$$q'_{i,l} = T_l(q_i), \qquad \text{intent}(q'_{i,l}) = \text{intent}(q_i) \tag{1}$$

Table 2 defines the four reporting families. The 14 labels and final counts are: arabic_script_confusion (120), spacing_variant (51), char_deletion (61), char_repetition (80), typing_transposition (54), homophone_spelling_confusion (5), homophone_or_typing_confusion (20), english_spacing_variant (70), english_char_deletion (40), english_transposition_or_repetition (40), english_punctuation_drop (10), question_variant (100), missing_qmark (60), and punctuation_drop (9). Family-level results are primary because several individual labels are small.

Table 2. Formal perturbation taxonomy.

Operational label	Family	Definition
Arabic-script confusion	Persian script/spacing	Arabic/Persian code-point substitution.
Spacing variant	Persian script/spacing	Join/split or whitespace/half-space change.
Character deletion	Persian character noise	Delete one Persian character.
Character repetition	Persian character noise	Repeat one Persian character.
Typing transposition	Persian character noise	Swap adjacent Persian characters.
Homophone spelling	Persian character noise	Plausible homophone/spelling substitution.
Homophone/typing	Persian character noise	Plausible question-word typing confusion.
English spacing	English-title noise	Insert/remove a space in an English title.
English deletion	English-title noise	Delete one English-title character.
English transposition/repetition	English-title noise	Swap or repeat an English-title character.
English punctuation drop	English-title noise	Remove punctuation from an English title.
Question variant	Query form/punctuation	Meaning-preserving reformulation.
Missing question mark	Query form/punctuation	Remove the final question mark.
Punctuation drop	Query form/punctuation	Remove query-level punctuation.

3.3.3. Construction and Quality Control

Perturbations were created through a controlled, query-aware process. Each variant inherited its parent's gold record and answer. Surface-identity checking first detected 39 intended perturbations that had remained unchanged; these were replaced with operator-consistent variants. A subsequent row-by-row human review corrected one bilingual title alignment across all six clean query types and

36 associated perturbations, and repaired 32 additional duplicated or operator-inconsistent query realizations. The final data retain 120 clean and 720 perturbed queries, six distinct variants per parent, the original query-type balance, and the original perturbation-label and family counts. The complete retrieval and answering pipeline was then rerun on the corrected benchmark and the 31,564-record cleaned corpus.

3.4. Retrieval Models

BM25 is computed over tokenized normalized profile text. Dense retrieval uses sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 with L2-normalized embeddings and cosine similarity. Character TF-IDF employs within-word-boundary 3-5-grams with sublinear term frequency. Hybrid retrieval combines min-max normalized BM25 and dense scores:

$$s_{\text{hybrid}}(d,q) = \alpha \cdot \tilde{s}_{\text{BM25}}(d,q) + (1-\alpha) \cdot \tilde{s}_{\text{dense}}(d,q) \qquad (2)$$

The primary system uses $\alpha = 0.5$, fixed a priori as an equal weighting of lexical and dense evidence before evaluation. The sweep over $\alpha \in \{0.0, 0.1, ..., 1.0\}$ is reported only as a post hoc sensitivity analysis on the clean and perturbed evaluation sets; it was not used to select or replace the primary configuration. The best observed evaluation performance occurs at $\alpha = 0.4$, whereas all principal results continue to use the prespecified $\alpha = 0.5$ setting.

3.5. Answering Models

The deterministic reference extracts the required field from the top-ranked hybrid record. The generative configuration uses the locally stored Qwen/Qwen2-7B-Instruct checkpoint [22] and receives the top five hybrid records. Its evidence context contains only query-relevant fields. Prompts require a year, director, Persian title, or the exact multi-field format year=<YEAR> | director=<DIRECTOR>. Decoding is greedy (do_sample=False, num_beams=1), with a 2,048-token input limit, task-dependent 12-80-token output limits, inference mode, and bfloat16 on GPU.

3.6. Evaluation and Statistical Analysis

Retrieval is evaluated using MRR@10 and Hit@k for k in {1, 3, 5, 10}; when the gold record is outside the top 10, its reciprocal rank is set to zero. Answering employs query-type-aware normalized field matching: years require exact normalized equality; director and Persian-title outputs are accepted when one normalized string contains the other; multi-field answers require all fields to satisfy their respective rules. The containment rule accommodates shortened or expanded surface forms but may accept overly short strings and is therefore reported as a limitation rather than being described as a strict exact match.

For incorrect outputs, retrieval-coverage failure means the gold record is absent from the top five; ranking/context-selection failure means it is present in the top five but not ranked first; and answer-stage failure means it is ranked first but the answer is incorrect. These mutually exclusive labels describe observable evidence states rather than proving a unique causal mechanism.

System differences are summarized using 10,000 paired bootstrap resamples and 95% percentile confidence intervals. Clean queries are resampled by query ID. For the perturbed benchmark, the parent query is treated as the sampling unit, and all six variants derived from that parent remain together within each resample. These variants share the same underlying information need, target movie, and gold answer and therefore cannot reasonably be treated as independent observations. Parent-level cluster resampling preserves this within-parent dependence, avoids pseudo-replication and artificially narrow intervals, and treats the 120 parent queries as the independent units of analysis. We report interval estimates and do not interpret the uncentered bootstrap distribution as a null-hypothesis p-value.

3.7. Reproducibility

The pipeline uses Pandas, Sentence-Transformers, rank-bm25, scikit-learn, and Hugging Face Transformers, with random seed 42 where randomness is used. Benchmark files, repair reports, retrieval rankings, prompts, generated answers, scored outputs, bootstrap results, and audit sheets are saved in one revision artifact directory.

4. Experimental Setup

The primary retrieval depth is 10. Both answering systems use rankings produced by the $\alpha = 0.5$ hybrid retriever, but the deterministic answer-extraction baseline accesses only the top-ranked record, whereas the RAG prompt contains the top five records. Consequently, the comparison varies both in evidence depth and answering mechanism. Any observed RAG advantage may therefore reflect access to additional candidate records, model-based evidence selection, generation, or a combination of these factors, and should not be interpreted as an isolated estimate of the contribution of generation.

5. Results

5.1. Retrieval Performance

Table 3 summarizes the overall retrieval performance on the clean and perturbed benchmarks. Hybrid retrieval achieves the strongest performance on both splits, reaching an MRR@10 of 81.50 on clean queries and 67.76 on perturbed queries. On the clean benchmark, this corresponds to improvements of 20.85 MRR@10 points over BM25, 29.11 points over dense retrieval, and 39.63 points over character n-gram TF-IDF. Under perturbation, the hybrid retriever maintains advantages of 17.54, 22.59, and 30.75 MRR@10 points over the same baselines, respectively. Although the character n-gram baseline exhibits the smallest absolute clean-to-perturbed decrease, its effectiveness remains substantially below that of the other retrievers on both splits. This indicates that partial character overlap alone is insufficient for reliably resolving bilingual movie records, even when it provides some tolerance to local spelling corruption.

Table 3. Overall retrieval performance (MRR@10, %).

System	Clean MRR@10	Perturbed MRR@10
BM25	60.65	50.21
Dense	52.40	45.17
Character n-gram	41.87	37.01
Hybrid ($\alpha = 0.5$)	81.50	67.76

Table 4 reports hybrid-retrieval performance separately for the six query types. The largest degradation occurs for English-title-to-Persian-title mapping, for which MRR@10 decreases from 71.67 to 51.19, corresponding to a drop of 20.48 points. Year-constrained title lookup is the second most affected query type, with a decline of 17.90 points. By contrast, title-to-director queries are comparatively stable, decreasing by only 6.83 points. These results show that query robustness depends strongly on the type of lexical evidence required for retrieval. Tasks that depend directly on corrupted English movie titles are more vulnerable than tasks whose Persian title or director cues remain relatively intact. This pattern is consistent with the perturbation-family analysis presented in Section 5.3.

Table 4. Hybrid retrieval by query type (MRR@10, %).

Query type	Clean	Pert.	Drop
Cross-lingual	88.21	73.10	15.11
EN -> FA title	71.67	51.19	20.48
Title -> director	82.50	75.67	6.83
Title -> year	81.00	68.86	12.14
Title -> year+director	80.62	70.61	10.01
Year-constrained	85.00	67.10	17.90

5.2. Answering Performance

Table 5 compares the top-1 deterministic extractive baseline with the top-5 field-specific RAG configuration. The extractive baseline achieves 75.83% accuracy on clean queries and 60.69% under perturbation. The RAG configuration improves these results to 84.17% and 72.08%, respectively. The resulting advantage of the RAG configuration is 8.33 percentage points on the clean benchmark and 11.39 points on the perturbed benchmark. The larger difference under perturbation suggests that access to multiple retrieved records, combined with model-based context selection and answer generation, can recover some answers that are unavailable to a rank-1-only extraction procedure. However, this comparison is not a controlled generation ablation. The deterministic answer-extraction baseline accesses only the top-ranked record, whereas the RAG model receives the top five retrieved records and performs model-based evidence selection and answer generation. Accordingly, the observed differences of 8.33 and 11.39 percentage points quantify the advantage of the complete evaluated top-5 RAG configuration over the top-1 deterministic answer-extraction configuration, rather than the independent effect of generation.

Table 5. Overall answering performance (%).

System	Clean	Perturbed	Drop
Top-1 extractive	75.83	60.69	15.14
Top-5 RAG	84.17	72.08	12.09

5.3. Robustness by Perturbation Family

Table 6 presents retrieval and answering performance for the four perturbation families. Query-form and punctuation variation is the least damaging family, with hybrid MRR@10 of 80.90, extractive accuracy of 75.74%, and RAG accuracy of 85.21%. Persian script and spacing variation is also comparatively well tolerated, producing hybrid MRR@10 of 75.49 and RAG accuracy of 77.78%. Performance decreases more noticeably for Persian character noise. In this family, hybrid MRR@10 falls to 71.36 and RAG accuracy to 71.36%. English-title noise is by far the most difficult condition, reducing hybrid MRR@10 to 40.65, extractive accuracy to 31.87%, and RAG accuracy to 53.12%. The results in Table 6 therefore show that the system is not uniformly sensitive to multilingual query variation.

Normalization appears to mitigate many Persian script and spacing differences, whereas deletion, spacing, transposition, and punctuation changes within short English movie titles often damage the

principal lexical evidence required for retrieval. Although the RAG configuration remains stronger than top-1 extraction in every family, it cannot fully compensate when the relevant movie record is not retrieved.

Table 6. Performance by perturbation family. Hybrid is evaluated using MRR@10, whereas the extractive and RAG systems are evaluated using accuracy.

Family	n	Hybrid MRR@10	Extractive	RAG
Query form	169	80.90	75.74	85.21
Persian script	171	75.49	69.01	77.78
Persian character	220	71.36	63.64	71.36
English title	160	40.65	31.87	53.12

5.4. Hybrid-Weight Sensitivity

Figure 1 illustrates the sensitivity of hybrid retrieval to the interpolation weight α. Performance improves as the system moves away from either dense-only retrieval at α = 0 or BM25-only retrieval at α = 1, confirming that the two score sources provide complementary information. Both clean and perturbed MRR@10 reach their maximum at α = 0.4. The best clean MRR@10 is 82.43, while the best perturbed MRR@10 is 68.87. The primary setting of α = 0.5 achieves 81.50 and 67.76, differing from the corresponding maxima by 0.93 and 1.12 points. Figure 1 therefore indicates that the principal hybrid-retrieval result is not dependent on a narrowly optimized coefficient. At the same time, α = 0.5 should be interpreted as a prespecified balanced setting rather than as the empirically optimal value.

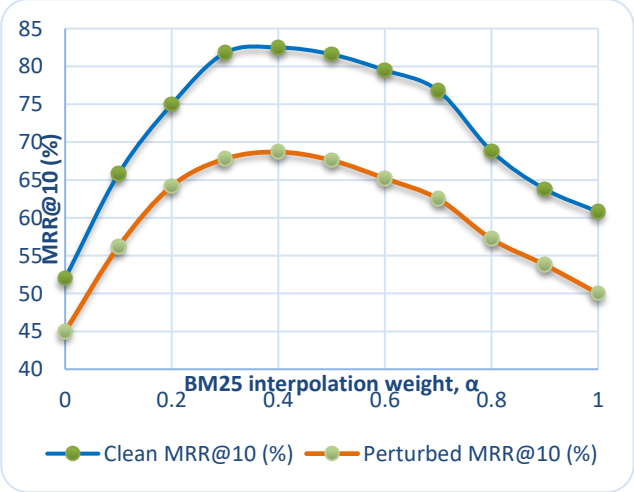

Figure 1. Hybrid retrieval MRR@10 across interpolation weights on clean and perturbed queries.

5.5. Confidence Intervals

Table 7 reports the observed paired system differences together with percentile 95% confidence intervals obtained from 10,000 bootstrap resamples. All reported intervals exclude zero, providing consistent evidence that the hybrid retriever outperforms BM25, dense retrieval, and character n-gram TF-IDF on both the clean and perturbed benchmarks. For retrieval, the hybrid improvement over BM25 is 20.85 MRR@10 points on clean queries, with a 95% confidence interval of 15.28–26.76, and 17.54 points on perturbed queries, with an interval of 13.76–21.47. The corresponding improvements over dense retrieval are 29.11 and 22.59 points, while the improvements over character n-gram retrieval are 39.63 and 30.75 points. Table 7 also shows that the top-5 RAG configuration exceeds top-1 extraction by 8.33 percentage points on clean queries, with a 95% confidence interval of 4.17–13.33, and by 11.39 points on perturbed queries, with an interval of 7.64–15.42. For the perturbed benchmark, bootstrap sampling is clustered by parent query. Consequently, the six perturbations derived from the same clean query remain together within each resample and are not treated as statistically independent observations.

Table 7. Paired retrieval differences in MRR@10 points and answering differences in accuracy points, with percentile 95% confidence intervals.

Comparison	Set	Δ	95% CI
Hybrid − BM25	Clean	20.85	15.28–26.76
Hybrid − BM25	Pert.	17.54	13.76–21.47
Hybrid − Dense	Clean	29.11	22.06–36.47
Hybrid − Dense	Pert.	22.59	17.25–28.16
Hybrid − Char.	Clean	39.63	33.24–46.06
Hybrid − Char.	Pert.	30.75	25.09–36.41
RAG − Extract.	Clean	8.33	4.17–13.33
RAG − Extract.	Pert.	11.39	7.64–15.42

All principal intervals exclude zero. The perturbed comparisons preserve parent-query clusters and therefore do not treat the six variants derived from one clean query as independent observations.

5.6. Failure Distribution and Consistency Audit

Table 8 summarizes the complete outcome distribution for the 720 perturbed queries evaluated with the RAG system. The model produces 519 correct answers, corresponding to an overall accuracy of 72.08%. Among the 201 incorrect outputs, 146 cases, or 20.28% of the complete benchmark, are classified as retrieval-coverage failures. In these cases, the gold movie is absent from the top-five retrieved records and is therefore unavailable to the answerer. A further 52 cases, representing 7.22% of the benchmark, are classified as ranking/context-selection failures. In these instances, the gold movie appears within the top-five context but is not ranked first. Among

these 52 cases, 48 involve an alternate corpus record sharing the same normalized Persian title and source QID as the query-specific gold record but differing in record-level metadata. Only three cases, corresponding to 0.42%, are classified as answer-stage failures, meaning that the gold record is ranked first but the final answer remains incorrect. Table 8 therefore indicates that retrieval coverage is the principal source of end-to-end failure, while exact record-level distinctions account for most ranking/context-selection errors. This conclusion is especially evident for English-title noise, which accounts for 73 retrieval-coverage failures, equivalent to 45.63% of that perturbation family. These results suggest that improving retrieval robustness is likely to provide greater benefit than modifying the generator alone. To verify implementation of the attribution rules, a stratified consistency audit was conducted on 43 errors. The audit included 20 retrieval-coverage cases, 20 ranking/context-selection cases, and all three available answer-stage cases. Each case was reviewed using the query, gold answer, model prediction, retrieved identifiers, and gold-record rank. All 43 reviewed cases were consistent with the deterministic rule-based label. This audit verifies implementation consistency but should not be interpreted as independent causal validation of the failure categories.

Table 8. Overall outcome distribution on the perturbed RAG benchmark.

Outcome	Count	Share (%)
Correct	519	72.08
Retrieval-coverage failure	146	20.28
Ranking/context-selection failure	52	7.22
Answer-stage failure	3	0.42

6. Discussion

The results support four principal conclusions. First, lexical and semantic retrieval signals are complementary. Hybrid retrieval consistently outperforms BM25, dense retrieval, and character n-gram TF-IDF on both clean and perturbed queries, while the interpolation analysis in Figure 1 shows that intermediate hybrid weights are preferable to either retrieval component alone. Second, evaluation on clean queries materially overestimates multilingual robustness. As shown in Tables 3–6, all retrieval and answering systems deteriorate under perturbation, but the magnitude of the decline varies substantially across query types and perturbation families. English-title corruption is particularly damaging, whereas query-form changes and Persian script or spacing variation are more effectively tolerated. Third, the evaluated top-5 RAG configuration outperforms top-1 deterministic answer extraction, particularly under perturbation. This improvement cannot be attributed exclusively to generation because the RAG model also receives a larger evidence set and can select among multiple candidate records. The result should therefore be interpreted as a comparison between two practical end-to-end configurations rather than as a controlled decomposition of generation quality. Fourth, the failure analysis in Table 8 shows that missing evidence is the dominant bottleneck. Most incorrect answers occur because the gold movie is absent from the top-five context, while answer-stage failures are rare. Most ranking and context-selection failures instead reflect competition between alternative records for the same underlying movie under the exact record-ID protocol. This indicates that stronger normalization, typo-robust retrieval, candidate expansion, reranking, and movie-level canonicalization may be more effective than prompt modification or the use of a larger generator alone.

7. Limitations

The benchmark contains 20 underlying bilingual movies and 120 clean query formulations. This matched design supports controlled paired analysis across query types and perturbations, but it is considerably smaller than broad multilingual retrieval benchmarks. The evaluation is also restricted to a single structured movie domain, one multilingual dense encoder, and one generator. Although BM25, dense retrieval, character n-gram TF-IDF, and hybrid retrieval represent several important retrieval families, the study does not include late-interaction or learned sparse models such as ColBERT and SPLADE. The perturbations are manually designed and controlled rather than collected from naturally occurring search logs. They therefore support precise diagnostic analysis but may not fully represent the frequency or distribution of real user errors. The answer-scoring procedure uses normalized field matching and permits containment matching for director and Persian-title strings. This accommodates shortened or expanded surface forms but may occasionally accept outputs that are too short. Evaluation also uses query-specific exact gold-record identifiers; alternate records for the same underlying movie are not accepted when their record-level metadata differ. Future evaluations should use curated alias sets, stricter entity-level answer matching, and canonical movie-level identifiers or explicit equivalence classes. Finally, the extractive and RAG configurations use different evidence depths.

The deterministic answer-extraction baseline accesses the top-ranked record, whereas the RAG model receives the top five records. Their comparison therefore does not isolate generation from additional context access and model-based evidence selection. Future work should conduct matched-depth experiments, including top-1 extraction versus top-1 RAG and top-5 extraction or deterministic selection versus top-5 RAG, so that the effects of evidence depth, evidence selection, and generation can be estimated separately.

8. Conclusion

This study introduced a controlled English-Persian movie-domain benchmark containing 120 clean queries and 720 meaning-preserving perturbations. Hybrid retrieval achieves 81.50 MRR@10 on clean queries and 67.76 under perturbation, while the evaluated top-5 field-specific Qwen2 configuration reaches 84.17% and 72.08% accuracy and outperforms top-1 deterministic answer extraction. The analysis also identifies substantial variation across perturbation conditions. Query-form and Persian script or spacing changes are comparatively well tolerated, whereas English-title corruption remains the primary weakness. Clustered bootstrap confidence intervals support the principal system comparisons, and the failure-distribution analysis shows that retrieval-coverage errors account for most incorrect RAG outputs.

These findings demonstrate that multilingual RAG robustness should be evaluated at multiple stages rather than through aggregate clean-input accuracy alone. Future work should enlarge the benchmark seed set, incorporate naturally occurring query errors, add domains and languages, adopt stricter alias-aware answer scoring, and evaluate normalization-aware, typo-robust, late-interaction, learned sparse, and reranking-based retrieval methods.

Data and Code Availability

The human-reviewed benchmark, evaluation code, prompts, scoring rules, retrieval outputs, generated answers, bootstrap summaries, and completed audit sheets will be released upon acceptance. Third-party corpus sources remain subject to their original licenses; the release will provide exact construction instructions and source identifiers.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, " in *Advances in Neural Information Processing Systems* 33, 2020, pp. 9459-9474.

[2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering, " in *Proc. EMNLP,* 2020, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.

[3] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktaschel, and S. Riedel, "KILT: A Benchmark for Knowledge Intensive Language Tasks, " in *Proc. NAACL-HLT,* 2021, pp. 2523-2544, doi: 10.18653/v1/2021.naacl-main.200.

[4] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation, " in *Proc. EACL System Demonstrations*, 2024, pp. 150-158, doi: 10.18653/v1/2024.eacl-demo.16.

[5] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems, " in *Proc. NAACL-HLT*, 2024, pp. 338-354, doi: 10.18653/v1/2024.naacl-long.20.

[6] R. Friel, M. Belyi, and A. Sanyal, "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems, " *arXiv:2407.11005, 2024*, doi: 10.48550/arXiv.2407.11005.

[7] P. Tasawong, W. Ponwitayarat, P. Limkonchotiwat, C. Udomcharoenchaikit, E. Chuangsuwanich, and S. Nutanong, "Typo-Robust Representation Learning for Dense Retrieval, " in *Proc. ACL,* vol. 2, 2023, pp. 1106-1115, doi: 10.18653/v1/2023.acl-short.95.

[8] G. Sidiropoulos and E. Kanoulas, "Improving the Robustness of Dense Retrievers Against Typos via Multi-Positive Contrastive Learning, " in *Advances in Information Retrieval: ECIR* 2024, Part III, Springer, 2024, pp. 297-305, doi: 10.1007/978-3-031-56063-7_21.

[9] N. Thakur, N. Reimers, A. Ruckle, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models, " in *Proc. NeurIPS Datasets and Benchmarks Track*, vol. 1, 2021.

[10] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, and J. Lin, "MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages, " *Trans. Assoc. Comput. Linguistics,* vol. 11, pp. 1114-1131, 2023, doi: 10.1162/tacl_a_00595.

[11] S. Percin, X. Su, Q. S. Syed, P. Howard, A. Kuvshinov, L. Schwinn, and K.-U. Scholl, “Investigating the Robustness of Retrieval-Augmented Generation at the Query Level, " in *Proc. Fourth Workshop on Generation, Evaluation and Metrics (GEM2)*, Vienna, Austria and virtual meeting, 2025, pp. 439-457, Association for Computational Linguistics.

[12] I. T. Sorodoc, L. F. R. Ribeiro, R. Blloshmi, C. Davis, and A. de Gispert, "GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation, " in *Findings of ACL*, 2025, pp. 17030-17049, doi: 10.18653/v1/2025.findings-acl.875.

[13] M. Song, S. H. Sim, R. Bhardwaj, H. L. Chieu, N. Majumder, and S. Poria, "Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse, " in *Proc. ICLR*, 2025, OpenReview: Iyrtb9EJBp.

[14] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, " in *Proc. SIGIR, 2020,* pp. 39-48, doi: 10.1145/3397271.3401075.

[15] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking, " in *Proc. SIGIR, 2021,* pp. 2288-2292, doi: 10.1145/3404835.3463098.

[16] M. A. C. Blandon, J. Talur, B. Charron, D. Liu, S. Mansour, and M. Federico, "MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation, " in *Proc. ACL,* 2025, pp. 22577-22595, doi: 10.18653/v1/2025.acl-long.1101.

[17] B. Li, F. Luo, S. Haider, A. Agashe, S. Li, R. Liu, M. M. Miao, S. Ramakrishnan, Y. Yuan, and C. Callison-Burch, "Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks: A Benchmark for Cross-lingual Robustness, " in *Findings of ACL, 2025*, pp. 4215-4241, doi: 10.18653/v1/2025.findings-acl.219.

[18] H. Hosseini, M. S. Zare, A. H. Mohammadi, A. Kazemi, Z. Zojaji, and M. A. Nematbakhsh, "PersianRAG: A Retrieval-Augmented Generation System for Persian Language, " in *Proc. 15th Int. Conf. Information and Knowledge Technology (IKT),* 2024, pp. 272-278, doi: 10.1109/IKT65497.2024.10892726.

[19] S. B. Hosseinbeigi, M. H. Shalchian, S. Asghari, M. A. Seif Kashani, and M. A. Abbasi, "Advancing Retrieval-Augmented Generation for Persian: Development of Language Models, Comprehensive Benchmarks, and Best Practices for Optimization, " in *Proc. 15th Language Resources and Evaluation Conference* (LREC 2026), 2026, pp. 7320-7330.

[20] A. Ghasemi, "Iranian Movies, " Kaggle dataset. [Online]. Available: https://www.kaggle.com/datasets/arianghasemi/iranian-movies. [Accessed: Jul. 11, 2026].

[21] R. Banik, "The Movies Dataset, " Kaggle dataset. [Online]. Available: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset. [Accessed: Jul. 11, 2026].

[22] A. Yang et al., "Qwen2 Technical Report, " *arXiv:2407.10671,* 2024, doi: 10.48550/arXiv.2407.10671.

پایداری سامانه‌های چندزبانه تولید تقویت‌شده با بازیابی (RAG) در برابر اغتشاش‌های پرس‌وجو: یک مجموعه‌معیار انگلیسی–فارسی

نیلوفر رنجبر[1,*] و حامد باغبانی[2]

[1] دانشکده فنی مهندسی جم، دانشگاه خلیج فارس بوشهر، جم، ایران.

[2] دانشکده فنی مهندسی جم، دانشگاه خلیج فارس بوشهر، جم، ایران.

ارسال ۲۰۲۶/۰۳/۱۴؛ بازنگری ۲۰۲۶/۰۷/۰۸؛ پذیرش ۲۰۲۶/۰۷/۱۵

چکیده:

سامانه‌های تولید تقویت‌شده با بازیابی (RAG) معمولاً با استفاده از ورودی‌های بدون خطا ارزیابی می‌شوند؛ در نتیجه، میزان پایداری آن‌ها در برابر تغییرات واقعی و تنوع‌های چندزبانه پرس‌وجوها به‌طور کامل مشخص نمی‌شود. در این پژوهش، یک مجموعه‌معیار پایداری انگلیسی–فارسی در حوزه فیلم برای ارزیابی عملکرد سامانه‌های RAG در شرایط تغییر و اغتشاش پرس‌وجو ارائه می‌شود. این مجموعه‌معیار بر پایه پیکره‌ای شامل ۳۱٬۵۶۴ رکورد، ۱۲۰ پرس‌وجوی اصلی و ۷۲۰ نمونه تغییریافته هم‌تراز ساخته شده است. مجموعه پیشنهادی، شش نوع قطعی پرس‌وجو و ۱۴ برچسب اغتشاش عملیاتی را که در چهار خانواده اصلی دسته‌بندی شده‌اند، پوشش می‌دهد. در این پژوهش، روش‌های BM25، بازیابی چگال چندزبانه، TF-IDF مبتنی بر n-گرم نویسه‌ای و بازیابی ترکیبی مقایسه شده‌اند. همچنین، عملکرد استخراج قطعی پاسخ از رتبه اول نتایج با یک سامانه RAG مبتنی بر Qwen2-7B-Instruct که از پنج نتیجه برتر بازیابی‌شده و اطلاعات مشخصه‌محور استفاده می‌کند، ارزیابی شده است. نتایج نشان می‌دهد که روش بازیابی ترکیبی در پرس‌وجوهای اصلی و اغتشاش‌یافته به‌ترتیب به امتیاز MRR@10 برابر با ۸۱٫۵۰ و ۶۷٫۷۶ دست یافته است. همچنین، سامانه RAG پیشنهادی به‌ترتیب دقت ۸۴٫۱۷ و ۷۲٫۰۸ درصد را در این دو شرایط کسب کرده است. بازه‌های اطمینان ۹۵ درصد حاصل از روش بوت‌استرپ جفت‌شده خوشه‌ای، معناداری تمامی تفاوت‌های اصلی بین سامانه‌ها را تأیید می‌کنند. تحلیل نتایج نشان می‌دهد که اغتشاش در عنوان‌های انگلیسی بیشترین اثر منفی را بر عملکرد سامانه دارد، در حالی که تغییرات مربوط به ساختار پرس‌وجو و نشانه‌گذاری در مقایسه با سایر تغییرات، بهتر تحمل می‌شوند. علاوه بر این، یک ارزیابی سازگاری شامل ۴۳ نمونه خطا، صحت پیاده‌سازی دسته‌بندی‌های مبتنی بر قواعد را تأیید می‌کند. بررسی کامل خروجی‌ها نیز نشان می‌دهد که خطاهای ناشی از عدم بازیابی رکورد مناسب، عامل اصلی شکست سامانه در شرایط دشوار، به‌ویژه در خانواده اغتشاش عنوان‌های انگلیسی، هستند. این یافته‌ها بر ضرورت ارزیابی چندمرحله‌ای و جزءبه‌جزء پایداری سامانه‌های RAG چندزبانه تأکید می‌کنند.

کلمات کلیدی: تولید تقویت‌شده با بازیابی؛ بازیابی اطلاعات چندزبانه؛ پایداری سامانه‌های RAG؛ بازیابی انگلیسی–فارسی؛ بازیابی ترکیبی؛ مجموعه‌معیار