# Raabta: Script-Aware Multi-Query Reformulation and Evidence Retrieval for Roman-Urdu Questions

Hasnat Khan
Assignment Project
Information Retrieval and Natural Language Processing

***Abstract*—Roman Urdu is widely used in digital communication, but it has no standard spelling and differs in script from most authoritative Urdu content. These properties make conventional lexical search brittle and can cause multilingual dense retrieval to return fluent-looking but unrelated passages. This paper presents Raabta, a CPU-first retrieval system that maps noisy Roman-Urdu questions to traceable Urdu evidence. The proposed pipeline preserves the original query, produces controlled normalized and Urdu-script views, retrieves with title-boosted BM25 and multilingual E5, adds character-level matching against romanized Urdu article titles, fuses complementary rankings, reranks a shallow candidate set, and applies relation-aware evidence checks before returning an extractive answer. Evaluation uses 120 project-verified development questions over 16,352 passages from a frozen 4,000-article Urdu Wikipedia subset; a separate 60-question test partition remains unused. Adding the romanized-title route raises Recall@10 from 0.1917 to 0.9833 and MRR@10 from 0.1014 to 0.5830 before reranking. The result is a transparent application that exposes every query view, retrieval route, score, validation gate, source, latency component, and abstention reason.**

***Index Terms*—Roman Urdu, Urdu, information retrieval, transliteration, multilingual embeddings, rank fusion, reranking, evidence grounding**

## I. Introduction

Roman Urdu expresses Urdu using the Latin alphabet. It is natural in messaging and social media, but it has no authoritative spelling standard: the same word may appear as *kya*, *kia*, or a shorter form, while vowels are frequently omitted from names. English insertions, abbreviations, and phonetic spelling add further variation. At the same time, reference material such as Urdu Wikipedia is written in Perso-Arabic script. A user can therefore express the correct intent while sharing few or no literal tokens with the relevant document.

This mismatch creates two related risks. First, a lexical system may retrieve nothing because the scripts differ. Second, a semantic retriever may find a vaguely associated passage and present it as if it answered the question. The second failure is particularly harmful in a question-answering interface because topical similarity is not the same as evidential support. Raabta treats the task as evidence retrieval rather than open-ended generation: it must locate a source, identify an exact supporting sentence, or abstain.

The research question is: *Can script-aware multi-query reformulation improve retrieval effectiveness for noisy Roman-Urdu questions over Urdu-script knowledge bases compared with direct retrieval, single transliteration, and standard hybrid retrieval?* The project makes four contributions:

1) a controlled query bridge that records original, normalized, Urdu-script, and retrieval-oriented views with accept/reject reasons;
2) a character 2–4-gram route that matches noisy Roman input directly against romanized Urdu article titles;
3) a multi-route retrieval and reranking pipeline with conservative source-alignment and relation-specific evidence gates; and
4) a reproducible CPU-only application and development evaluation whose interface exposes the complete decision path.

The paper distinguishes retrieval recall from final answer accuracy. The main reported improvement is Recall@10 on a title-oriented development set. It indicates whether relevant evidence enters the first ten candidates; it is not a claim that the system correctly answers 98.33% of arbitrary questions.

## II. Related Work

This section summarizes the five papers reviewed for the project and identifies the gap addressed by Raabta.

### A. Roman-Urdu Resources and Retrieval

Alam and Hussain introduced Roman-Urdu-Parl, a large parallel Roman-Urdu/Urdu corpus assembled from multiple sources and quality-checked through crowd processes [1]. It provides valuable script-aligned evidence, but a parallel corpus does not itself define document relevance or an end-to-end retrieval protocol. Raabta uses a bounded slice only to support transliteration analysis; it never treats parallel sentences as relevance labels.

Butt, Varanasi, and Neumann studied transformer-based transliteration between Roman Urdu and Urdu [2]. Their work demonstrates the value of multilingual transfer and stronger character-level evaluation for an orthographically variable language. However, selecting one transliteration remains risky for search because several spellings can encode the same entity. Raabta retains multiple controlled query views and adds a title route rather than requiring a single conversion to be correct.

The same authors introduced a large-scale Roman-Urdu information-retrieval dataset and a multilingual retrieval baseline [3]. That work establishes that trained multilingual retrieval can outperform lexical baselines on a translated MS MARCO-derived benchmark. The remaining gap is a transparent, CPU-first system that searches a native Urdu-script knowledge collection, exposes each query transformation, measures components separately, and refuses unsupported evidence.

### B. Multilingual Retrieval and Reranking

Wang *et al.* present multilingual E5 encoders trained through multilingual contrastive learning [4]. E5 provides a practical shared embedding space for cross-script semantic retrieval, including a small 384-dimensional model suitable for CPU inference. Dense similarity nevertheless remains vulnerable to fine-grained entity and relation errors. RAABTA therefore combines E5 with lexical and character-level title evidence instead of using dense scores alone.

Nogueira and Cho show that BERT-style cross-encoders can substantially improve passage reranking by jointly encoding a query and candidate [5]. Their result motivates the final reranking stage, but cross-encoders are computationally expensive on CPU and cannot guarantee that a high-scoring passage contains the exact requested fact. RAABTA reranks only 20 fused candidates and follows the model score with deterministic evidence checks.

Together, these five studies provide parallel data, transliteration, a Roman-Urdu retrieval benchmark, multilingual embeddings, and neural reranking. They do not jointly solve noisy entity matching, calibrated evidence refusal, route-level transparency, and CPU deployment over a fixed Urdu corpus. This intersection is the focus of the proposed system. BM25 [6] and reciprocal rank fusion (RRF) [7] provide the complementary classical retrieval and score-independent fusion foundations used below.

## III. DATASET

### A. Sources

The primary knowledge source is the Urdu configuration of the Wikimedia Wikipedia dataset, snapshot `20231101.ur` [8]. A deterministic selection procedure retains 4,000 articles across history, geography, science, culture, Pakistan, and general topics. Each record preserves its article identifier, title, URL, raw text, normalized text, and assigned domain. The subset is deliberately bounded so that experiments remain reproducible and can run on a normal CPU.

Roman-Urdu-Parl supplies a fixed 30,000-row supporting slice for spelling and transliteration analysis [1]. It does not provide gold retrieval passages and is not used to train or evaluate relevance. This separation prevents the auxiliary parallel resource from leaking target article or passage information into query generation.

### B. Preprocessing

Text is normalized with Unicode NFKC while both raw and normalized forms remain available for audit. Articles shorter than 80 tokens are excluded by the frozen preprocessing configuration. The default chunker creates passages of at most 150 whitespace tokens with 30-token overlap, producing 16,352 passages. Each passage carries a deterministic identifier, article identifier, title, text span, URL, domain, token count, and checksum-linked provenance. The default passage artifact has SHA-256 `47648cf6...2e5671`; the complete value is stored in project metadata.

TABLE I
DATASET AND EVALUATION SUMMARY

| Item | Value |
|---|---|
| Urdu Wikipedia articles | 4,000 |
| Default passages | 16,352 |
| Passage length / overlap | 150 / 30 tokens |
| Supporting parallel rows | 30,000 |
| Diagnostic questions | 180 |
| Development / locked test | 120 / 60 |
| Domains | 6 |
| Query categories | 8 |

The diagnostic set contains 180 evidence-linked, title-definition questions. It spans clean Roman Urdu, informal spelling, heavy noise, abbreviations, Urdu-English code switching, named entities, short queries, and slight ambiguity. Each record stores the Roman query, Urdu form, gold passage identifier, exact evidence, query type, domain, split, and review status. The split was frozen at 120 development and 60 test questions before model tuning. Only the development partition is used in this paper.

## IV. METHODOLOGY

Figure 1 summarizes the end-to-end architecture. All components use pinned versions and local artifacts.

### A. Controlled Query Views

QUERYBRIDGE preserves the original query and can create three additional views: a conservatively normalized Roman form, an Urdu-script transliteration, and a controlled retrieval-oriented form. The generator records the source query, transformation type, semantic similarity, transliteration coverage, and decision reason. Exact duplicates and low-similarity variants are rejected. Query generation never receives the gold article, gold passage, answer, or evidence text.

### B. Three First-Stage Retrieval Routes

The lexical route uses Unicode-aware BM25 with article-title boosting. BM25 provides interpretable exact-term evidence and follows the probabilistic relevance framework [6]. A zero-overlap safeguard prevents arbitrary documents with zero lexical score from entering fusion.

The semantic route uses `intfloat/multilingual-e5-small`, pinned to revision `d1d99a1...70e51`. Passage embeddings are generated offline with the E5 passage prefix and stored as a $16{,}352 \times 384$ float32 matrix. Queries use the corresponding query prefix and

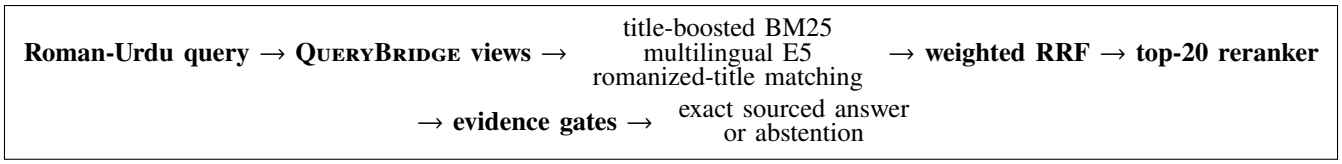


Fig. 1. The proposed script-aware retrieval and evidence-validation pipeline. The optional live Urdu Wikipedia route is queried only after local abstention and explicit user consent.

exact normalized dot-product search. Sentence-Transformers provides the local model interface [9].

The proposed title route addresses the main observed weakness: named entities written noisily in Roman characters. One lead passage is retained for each article, and its Urdu title is romanized with `uroman` [10]. Common question boilerplate is removed from the input. Scikit-learn [11] then builds character-boundary TF-IDF features over 2–4-grams. This representation tolerates vowel insertion or deletion: a consonant-heavy romanization can still match a longer informal spelling. Restricting the route to one lead passage per article prevents duplicate chunks from dominating the list.

### C. Fusion and Reranking

Every accepted query view traverses BM25 and E5, while the original query also enters the title route. Weighted RRF combines ranks without assuming that lexical, dense, and character scores are calibrated on a common scale:

$$S(d) = \sum_{r \in R} \frac{w_r}{k + \mathrm{rank}_r(d)}, \tag{1}$$

where $k = 60$ and $w_r$ is a frozen route weight. RRF is suitable because it is training-free and robust to heterogeneous score ranges [7]. The romanized-title route receives the largest weight (3.0), Urdu-script BM25 1.35, and other view/route combinations between 0.55 and 1.20.

The best 20 fused passages are rescored by the pinned `gte-multilingual-reranker-base` cross-encoder [12]. Each pair concatenates the article title and passage text so that an exact title match remains visible to the reranker. This improves precision but is the largest CPU cost.

### D. Grounded Evidence and Interface

A top passage must clear a 0.62 reranker relevance threshold and show meaningful converted-term overlap or cross-script title alignment. Candidate sentences are compared with accepted query views and must exceed 0.70 semantic similarity. Relation-aware rules then check the expected answer shape: birth and death questions require the correct temporal relation; prices require a currency and amount; quantity questions require a number; and current-capital questions reject historical statements. Navigation text, references, and generic list descriptions are excluded. Extraction is restricted to the exact source that passed validation.

The FastAPI backend [13] returns structured stages, route counts, accepted and rejected variants, scores, thresholds, sources, candidate evidence, decisions, and latency. The React interface [14] displays this trace in plain language. Live Urdu Wikipedia search is off by default and can run only when the local pipeline abstains and the user explicitly enables it; live passages must pass the same validation gates.

## V. Experimental Setup

### A. Systems and Metrics

The frozen comparison includes: (1) direct dense retrieval; (2) one deterministic transliteration followed by BM25; (3) a conventional raw-query BM25+dense RRF hybrid; (4) QueryBridge with BM25, dense retrieval, and RRF; (5) the same system with depth-20 reranking; and (6) the current retrieval pipeline with the romanized-title route.

Retrieval is measured with Recall@1, Recall@5, Recall@10, MRR@10, and nDCG@10. Recall@10 is the primary coverage measure, MRR@10 reflects how early the first relevant passage appears, and nDCG@10 rewards useful ordering across the top ten. Metrics are macro-averaged over 120 development questions. No test query is evaluated.

### B. Validation and Reproducibility

The development/test split is stored in the diagnostic CSV rather than regenerated. Model names, revisions, seeds, passage settings, route limits, and the RRF constant are frozen in YAML and metadata. Passage and embedding checksums are verified before evaluation. Retrieval ablations independently disable normalization, transliteration, expansion, BM25, dense retrieval, or fusion. Robustness is grouped by eight frozen query categories. Latency excludes cold model loading unless explicitly stated.

All Python dependencies are pinned, seven notebooks are executed with visible outputs, and 47 automated tests cover preprocessing, query generation, retrieval, reranking, evidence checks, backend responses, and the Wikipedia fallback. A portability audit verifies Python support, exact dependency pins, artifact checksums, notebook execution, frontend build presence, relative paths, and zero test-set usage in evaluation reports.

### C. Evaluation Protocol

Each diagnostic question has one verified gold passage identifier. For a ranked list $L_q$ and cut-off $k$, Recall@$k$ is one when that identifier occurs in the first $k$ results and zero otherwise. MRR@10 is the reciprocal rank of the first relevant passage, or zero when it is absent; nDCG@10 uses the same binary relevance judgment with logarithmic rank discount. Reported scores are arithmetic means over the 120 development questions. This design measures whether the

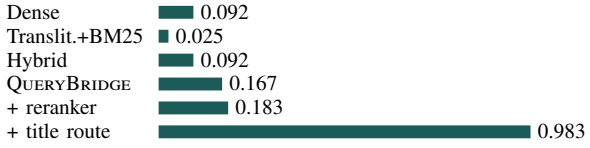


Fig. 2. Recall@10 by retrieval system on the development split.

retrieval stage supplies the answer layer with the expected evidence. It does not credit a topically similar passage, nor does it evaluate fluency.

Every compared system receives the same corpus, passage boundaries, query order, and relevance labels. Baselines differ only in the stated query transformation and retrieval components. The title-route regression uses the same development questions and the same no-reranker setting before and after the route is enabled, which isolates its contribution more clearly than comparing it with the slower reranked configuration. The fixed test flag is checked across all generated evaluation reports, and its required value is zero. Consequently, hyperparameters and failure rules remain development decisions; the final test result is deliberately not reported.

Reproduction starts from the pinned Python environment and the frozen corpus metadata. The passage checksum is `47648cf6...2e5671`, and the 16,352 × 384 embedding array checksum is `80a612c0...83bc9a`. The seven analysis notebooks correspond to data setup, exploratory analysis, baselines, QueryBridge, ablation, robustness, and error analysis. Machine-specific absolute paths are excluded. These controls make a transferred project verifiable without claiming bit-identical timing across processors.

## VI. Results

### A. Baseline Comparison

Table II reports the original system comparison. Direct dense and standard hybrid retrieval both reach 0.0917 Recall@10. The single-transliteration baseline is weaker, confirming that one brittle conversion is insufficient. Multi-view QueryBridge raises Recall@10 to 0.1667. Reranking improves early ordering (MRR@10 from 0.0750 to 0.1444) but adds substantial CPU latency.

### B. Proposed Title-Route Improvement

The application regression in Table III compares the preceding local retrieval pipeline against the same pipeline after adding romanized-title matching. Recall@10 rises from 0.1917 to 0.9833, an absolute increase of 0.7916. Recall@5 rises by 0.7583 and Recall@1 by 0.3167. MRR@10 improves from 0.1014 to 0.5830. Figure 2 shows the size of the Recall@10 change relative to the earlier systems.

### C. Ablations, Robustness, and Cost

Table IV reports retrieval-stage leave-one-component-out controls before the title-route addition. Removing fusion produces the largest observed loss, followed by transliteration and dense retrieval. Removing expansion slightly lowers Recall@10 but raises MRR@10, so expansion is a mixed component rather than a universal improvement.

The original reranked pipeline varied sharply by query category: clean queries reached 0.5000 Recall@10, while abbreviated queries reached 0.0000. Highly noisy and short queries reached 0.2667, code-switched queries 0.2143, and named-entity queries only 0.0714. This failure pattern directly motivated character-level title retrieval. A post-change category-wise evaluation has not been substituted for the original robustness table; the current 0.9833 result is reported only as an aggregate regression on the same questions.

The 16,352 × 384 embedding matrix occupies 25.1 MB and required 42.4 minutes for cold CPU generation and serialization. Direct dense retrieval averages 51.8 ms, standard hybrid 94.5 ms, and the original QueryBridge retrieval 444.0 ms. The title-route regression averaged approximately 580.8 ms before reranking. Depth-20 reranking adds about 16.4 s and raises observed resident memory to approximately 2.49 GB, making it the dominant interactive cost.

### D. Trace-Level Analysis

The grounded-QA smoke case asks *"pakistan ka capital kya hai"*. QueryBridge retains the original, rejects an identical normalized form, and accepts both an Urdu-script view and a retrieval-oriented view. The fused retrieval places the article "Pakistan ke Dar-ul-Hakoomat" at rank one. Evidence selection returns the source's two lead sentences, including the explicit statement that Islamabad has been the national or federal capital since 1960. Sentence similarities are 0.8783 and 0.8449, both above the 0.70 evidence threshold. Retrieval takes 381.5 ms and answer selection 1,396.9 ms in this single local run. This is a worked trace, not an aggregate latency or accuracy claim.

The example illustrates why the interface exposes intermediate state. A user can see that the duplicate query was rejected, which script-converted views were searched, which routes contributed to the fused candidate, why the sentence passed the current-capital rule, and which URL supports the answer. If the relevant article had been retrieved but the sentence lacked a current-capital statement, the trace would show a relation-gate rejection rather than silently presenting the passage. If no source crossed the retrieval and evidence thresholds, the response would contain an abstention reason and no fabricated citation.

## VII. Discussion

### A. Interpretation

The strongest result is not produced by a larger generative model. It comes from matching the structure of the problem: Roman-Urdu entity spellings are noisy, while Urdu Wikipedia titles provide concise article identities. Romanizing those titles once and comparing them with character *n*-grams creates a high-recall bridge that is robust to missing vowels and short spelling variations. The lead-passage restriction then converts

TABLE II
DEVELOPMENT RETRIEVAL RESULTS ($n$ = 120; LOCKED TEST QUERIES USED = 0)

| System | R@1 | R@5 | R@10 | MRR@10 | nDCG@10 | Mean ms |
|---|---|---|---|---|---|---|
| Direct dense | 0.0500 | 0.0833 | 0.0917 | 0.0615 | 0.0686 | 51.8 |
| Single transliteration + BM25 | 0.0000 | 0.0167 | 0.0250 | 0.0074 | 0.0116 | 60.8 |
| Standard hybrid | 0.0250 | 0.0667 | 0.0917 | 0.0456 | 0.0565 | 94.5 |
| QUERYBRIDGE + RRF | 0.0500 | 0.1000 | 0.1667 | 0.0750 | 0.0959 | 444.0 |
| QUERYBRIDGE + depth-20 reranker | 0.1250 | 0.1750 | 0.1833 | 0.1444 | 0.1540 | 16,947.6 |

TABLE III
APPLICATION RETRIEVAL REGRESSION

| Metric | Before | With title route |
|---|---|---|
| Recall@1 | 0.0750 | **0.3917** |
| Recall@5 | 0.1167 | **0.8750** |
| Recall@10 | 0.1917 | **0.9833** |
| MRR@10 | 0.1014 | **0.5830** |
| nDCG@10 | 0.1219 | **0.6796** |

TABLE IV
RETRIEVAL ABLATION RESULTS

| Configuration | R@10 | MRR@10 | Mean ms |
|---|---|---|---|
| Full, no reranker | 0.1667 | 0.0750 | 425.2 |
| No normalization | 0.1500 | 0.0715 | 402.9 |
| No transliteration | 0.0667 | 0.0253 | 335.8 |
| No expansion | 0.1583 | 0.0843 | 354.4 |
| No BM25 | 0.1417 | 0.0530 | 271.1 |
| No dense | 0.0833 | 0.0488 | 254.8 |
| No fusion | 0.0583 | 0.0232 | 489.2 |

TABLE V
ORIGINAL 30-CASE FAILURE AUDIT

| Assigned failure category | Cases |
|---|---|
| Roman spelling mismatch | 5 |
| Excessive spelling noise | 5 |
| Code-switching failure | 5 |
| Named-entity mismatch | 5 |
| Short or ambiguous query | 5 |
| Irrelevant retrieval | 5 |

a title match into a useful evidence candidate without flooding fusion with multiple chunks from one article.

The earlier ablation findings remain informative. Transliteration helps cross the script boundary, dense retrieval supplies semantic coverage, and RRF prevents one score family from dominating. The mixed expansion result warns against assuming that every additional query view is beneficial. The title route succeeds because it targets a measured failure category rather than adding unconstrained paraphrases.

Reranking and evidence validation solve different problems. The cross-encoder reorders candidates but does not prove that the selected sentence answers the requested relation. Deterministic gates therefore reject a birth answer without birth wording, a price without currency and amount, and a historical capital statement when the user asks for the current capital. This separation also improves transparency: the interface can state whether failure occurred during retrieval, source alignment, sentence selection, or relation validation.

### *B. Failure Cases and Limitations*

A stratified 30-case audit of the original pipeline assigns five cases to each category in Table V. The audit is diagnostic rather than an estimate of naturally occurring error prevalence. The final system directly addresses spelling and entity errors through title matching and improves irrelevant-answer behavior through source and relation gates. Code switching remains difficult when an English token names an entity that is represented differently in the Urdu title. Very short or ambiguous questions also remain difficult because a responsible system should request clarification rather than infer unstated intent.

Four limitations constrain the conclusions. First, the diagnostic set is title-oriented, so it favors tasks where article identity is a strong signal. Second, the local corpus contains only 4,000 articles and cannot represent all Urdu information needs. Third, the 120-question development set is small and the 60-question test set remains intentionally unused. Fourth, the current regression measures retrieval rather than independently reviewed end-to-end answer correctness. The 0.9833 Recall@10 score must not be read as a general answer-accuracy percentage.

There are also methodological threats to validity. Question construction may share vocabulary patterns with the source titles even though gold fields are unavailable to query generation. One relevant passage per question simplifies scoring and may under-credit alternative passages that contain equally valid evidence. Thresholds were tuned through development analysis and are not probabilistically calibrated. The failure labels were assigned by rules and require independent human review. Finally, CPU timings reflect one local environment and should be compared by order of magnitude rather than treated as hardware-independent constants.

The optional live Wikipedia route improves encyclopedic coverage but introduces network dependence and cannot reliably answer current shopping prices or breaking news. The interface therefore keeps it off by default, discloses when a query leaves the PC, and subjects live passages to the same evidence checks. Dataset attribution and Wikipedia licensing must be preserved whenever processed text is redistributed.

## VIII. Conclusion and Future Work

This paper presented Raabta, a complete CPU-first system for connecting noisy Roman-Urdu questions to Urdu-script evidence. Controlled query views, BM25, multilingual E5, weighted RRF, title-aware reranking, and extractive validation form an auditable retrieval pipeline. The proposed romanized-title route produces the largest measured improvement, raising development Recall@10 from 0.1917 to 0.9833 and MRR@10 from 0.1014 to 0.5830 before reranking. The application complements this retrieval gain with relation-aware evidence checks and visible abstention reasons, reducing the chance that a semantically related but unsupported passage is presented as an answer.

Future work should first obtain independent language review of query naturalness, gold evidence, and failure labels, then freeze the configuration and run the locked test partition once. A category-wise evaluation of the title route would show whether gains are concentrated in named entities or extend to abbreviations and code switching. Character encoders or learned alias models could replace fixed TF-IDF title matching, while calibrated confidence models could combine title, reranker, and evidence scores. A shallower or distilled reranker is needed for interactive CPU latency. Finally, broader locally licensed Urdu collections and an explicit clarification dialog would improve coverage without weakening the system's evidence-first behavior.

## References


[1] M. Alam and S. U. Hussain, "Roman-urdu-parl: Roman-urdu and urdu parallel corpus for urdu language understanding," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–20, 2022.

[2] U. Butt, S. Varanasi, and G. Neumann, "Low-resource transliteration for roman-urdu and urdu using transformer-based models," in *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages*, 2025, pp. 144–153. [Online]. Available: https://aclanthology.org/2025.loresmt-1.13/

[3] M. U. T. Butt, S. Varanasi, and G. Neumann, "Roman urdu as a low-resource language: Building the first ir dataset and baseline," in *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, 2025, pp. 82–87. [Online]. Available: https://aclanthology.org/2025.lowresnlp-1.9/

[4] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," *arXiv preprint arXiv:2402.05672*, 2024. [Online]. Available: https://arxiv.org/abs/2402.05672

[5] R. Nogueira and K. Cho, "Passage re-ranking with bert," *arXiv preprint arXiv:1901.04085*, 2019. [Online]. Available: https://arxiv.org/abs/1901.04085

[6] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[7] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd International ACM SIGIR Conference*, 2009, pp. 758–759.

[8] Wikimedia Foundation, "Wikipedia dataset: Urdu configuration 20231101.ur," 2023, frozen revision 3e1f92c331f318af862b87e2319ed5dc26d80f5d. [Online]. Available: https://huggingface.co/datasets/wikimedia/wikipedia

[9] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[10] USC Information Sciences Institute, "uroman: Universal romanizer, version 1.3.1.1," 2026. [Online]. Available: https://github.com/isi-nlp/uroman

[11] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] Alibaba NLP, "gte-multilingual-reranker-base model card," 2024. [Online]. Available: https://huggingface.co/Alibaba-NLP/gte-multilingual-reranker-base

[13] S. Ramírez, "Fastapi documentation," 2026. [Online]. Available: https://fastapi.tiangolo.com/

[14] Meta Open Source, "React documentation," 2026. [Online]. Available: https://react.dev/