A Unique Ranking Framework (URF) with LLM-Based Reranking for English-Hindi Bilingual Corpus Search

Sanjog Arora[1,2], Naveen Hemrajani[1,*]

[1] JECRC University, Plot No. IS-2036 to 2039, Ramchandrapura Industrial Area, Vidhani, Sitapura Extension, Jaipur 303905, Rajasthan, India

[2] Anand International College of Engineering, Near Kanota, Agra Road, Jaipur 303012, Rajasthan, India

*Author for correspondence: dean.engineering@jecrcu.edu.in

Abstract

Cross-lingual information retrieval (CLIR) allows users to retrieve relevant information written in a language different from that of the query, and therefore plays a critical role in multilingual knowledge access. Recent advances in dense retrieval and retrieval-augmented generation (RAG) have improved semantic matching; however, current retrieval schemes do not fully align semantic understanding with lexical matching, particularly for morphologically diverse and low-resource language pairs such as Hindi and English. In addition, conventional ranking fusion strategies depend primarily on rank positions and often fail to exploit the underlying semantic and lexical scores.

The present investigation proposes a Unique Ranking Framework (URF) that integrates dense semantic retrieval, BM25-based lexical retrieval and large language model (LLM) based reranking within a combined RAG architecture. A new bilingual benchmark, BiIR-2026, consisting of 1,860 aligned English-Hindi (En-Hi) document pairs and 742 relevance-judged queries, was constructed from publicly available resources. Experimental evaluation demonstrated that the proposed URF outperformed representative dense, lexical, sparse-neural and late-interaction retrieval baselines across Precision@10, Recall@10, mean average precision (MAP), mean reciprocal rank (MRR) and normalized discounted cumulative gain at rank 10 (NDCG@10). Ablation studies further showed that optimised score fusion and selective LLM reranking improved retrieval effectiveness while keeping computational latency feasible for interactive platforms. The proposed framework provides a reproducible and scalable alternative for multilingual retrieval and offers a practical basis for extending RAG to low-resource Indic languages.

Keywords: cross-lingual information retrieval (CLIR), BM25, retrieval-augmented generation (RAG), large language models (LLM), semantic retrieval, hybrid information retrieval, multilingual retrieval, English-Hindi retrieval, vector search, FAISS

Introduction

The volume of digitised bilingual and multilingual content produced by government, educational, legal and journalistic institutions in India has increased considerably, yet the tools and techniques used to search this content remain largely monolingual (Jagarlamudi & Kumaran, 2007). A person searching in Hindi often needs information held in an English document, and translating such documents in advance is neither feasible nor affordable (Katta

& Arora, 2015). Cross-lingual information retrieval (CLIR) offers a feasible and efficient solution to this problem by enabling a query written in one language to return relevant results in another, without prior translation effort (Dwivedi & Chandra, 2016; Goworek et al., 2025; Zhou et al., 2012).

In cross-lingual and multilingual information retrieval for Indian languages, dictionary-based query translation has been widely used, in which the query is translated word by word using a bilingual lexicon and matched against a monolingual index (Dave & Majumder, 2026; Dwivedi & Chandra, 2016; Seetha et al., 2007). Dictionary-based CLIR is easy to construct and requires no parallel training data (Levow et al., 2005; Nic Gearailt, 2005); however, translation ambiguity, out-of-vocabulary words and the loss of multi-word expression semantics limit its retrieval effectiveness (Hedlund et al., 2004; Sharma & Mittal, 2016).

Over the last few years, query translation quality has improved considerably, and dual-encoder sentence embedding models such as Language-agnostic BERT Sentence Embedding (LaBSE) have transformed CLIR by placing semantically similar sentences from different languages in a common vector space (Feng et al., 2022; Nawaz, 2025). This allows direct cross-lingual matching without explicit translation, which in turn improves both the efficiency and the accuracy of multilingual retrieval systems (Rahimi et al., 2015). Resource scarcity remains a key limitation for Indian languages, prompting the adoption of transfer-learning techniques that adapt English-centric retrieval models with minimal fine-tuning on Indic-language data (Bhattacharya et al., 2018; Puranegedara et al., 2025).

Sparse retrieval approaches that rely on word frequency, including term frequency (TF), inverse document frequency (IDF) and the Okapi BM25 probabilistic ranking function, have provided the foundation of practical, efficient and robust search engines, particularly where query and document vocabularies overlap (Iswarya & Radha, 2012; Jones et al., 2000; Robertson & Zaragoza, 2009). Learning-to-rank (LTR) approaches subsequently extended these manually designed scoring methods by framing ranking as a supervised learning problem: RankNet and RankSVM order document pairs efficiently, while LambdaMART directly optimises rank-sensitive metrics such as normalized discounted cumulative gain (NDCG) (Chavhan & Dharmik, 2025; Zhu & Klabjan, 2020). These approaches are established benchmarks in retrieval-augmented generation (RAG) because they integrate manually crafted attributes such as term overlap, proximity and document length normalisation into a unified learned scoring mechanism. The present study extends this line of work by deriving a fusion weight from both neural and lexical inputs rather than from lexical characteristics alone.

Recent advances in RAG have made it possible to combine search over documents stored in a vector database with a generator, that is, a large language model (LLM), which produces a precise natural-language answer derived from the retrieved documents (Arslan et al., 2024; Lewis et al., 2020). RAG couples a retriever with a generative language model and mitigates the hallucination and knowledge-cutoff problems of standalone LLMs by conditioning generation on retrieved facts; its retriever component can be made language-agnostic by using multilingual sentence embeddings (Oche et al., 2025; Shahzad et al., 2026; Zhang & Zhang, 2025). However, semantic retrieval based on dense embeddings usually underperforms on queries

containing names, numbers or rare terms, where lexical overlap carries most of the discriminative signal (Hambarde & Proenca, 2023; Sciavolino et al., 2021).

Lexical methods such as BM25, conversely, fail when the query and the relevant passage share almost no surface tokens (Gao et al., 2021; Robertson & Zaragoza, 2009; Wang et al., 2024), a common situation in cross-lingual search. This has driven the development of hybrid retrieval, which merges dense semantic search with lexical search and increasingly introduces a third stage in which LLMs perform reranking (Luan et al., 2021; Wang et al., 2024). Dense passage retrieval (DPR) has further shown that a contrastively trained dual encoder can outperform BM25 on open-domain queries by capturing semantic similarity (Karpukhin et al., 2020). LLMs have been reported to deliver better ranking quality on paraphrase, negation and implication than a combination of vector and bag-of-words systems (Sun et al., 2023; Zhu et al., 2025); however, the integration of semantic similarity, lexical similarity and LLM judgement into a single flexible and economical ranking framework has not been reported or evaluated for cross-lingual English-Hindi retrieval with two-stage, corpus-tunable weighting (Hsu & Tzeng, 2025; Rao et al., 2025). In addition, existing multilingual RAG systems generally rely on a single fusion approach such as reciprocal rank fusion (RRF) (Kafi et al., 2026; Rajaei et al., 2024), and few studies have isolated the specific contribution of an LLM reranking stage in a bilingual Indian-language setting (Nair & Gupta, 2026). Moreover, general multilingual embedding models are trained predominantly on high-resource language pairs and underperform on English-Hindi retrieval for conceptual queries (Sree et al., 2026). An adjustable hybrid fusion approach is therefore needed, one that can be tuned to a corpus rather than fixed a priori, together with a reproducible bilingual evaluation set carrying graded relevance judgements rather than binary success or failure indicators.

This study presents the design, implementation and evaluation of a cross-lingual retrieval framework, the Unique Ranking Framework (URF), which integrates semantic search, lexical search and LLM-based cross-encoder reranking through a two-stage scoring function governed by learned mixing parameters. A curated English-Hindi evaluation corpus (BiIR-2026) of 1,860 query-document pairs drawn from Press Information Bureau (PIB) press releases, the Samanantar corpus and comparable Wikipedia articles was developed to evaluate factoid, keyword and conceptual queries annotated with graded relevance judgements.

Methodology

Bilingual Corpus and Data Preparation

Corpus characteristics

In the present study a new benchmark, BiIR-2026, was constructed rather than reusing an existing monolingual IR test collection, because the available relevance benchmarks do not target English-Hindi document-level retrieval across mixed domains. The corpus was assembled from three sources, chosen to represent realistic bilingual search demand in the Indian context:

(i) a document-aligned and filtered subset of the Samanantar parallel corpus, restricted to aligned sentence groups;

(ii) bilingual press releases published by the Press Information Bureau (PIB), Government of India;

(iii) comparable English-Hindi Wikipedia article pairs identified through inter-language links.

After a duplicate-removal step that discarded documents with a cosine similarity above 0.97, and length filtering that discarded documents shorter than 150 words or longer than 8,000 words, the benchmark comprised 1,860 aligned English-Hindi document pairs (3,720 documents in total; Table 1). Segmentation of the corpus yielded 14,275 indexed passages, from which 742 evaluation queries were constructed. Each query was associated with graded relevance judgements (0 = not relevant, 1 = marginally relevant, 2 = relevant, 3 = highly relevant) for the top 20 passages retrieved by a pooling procedure across all baseline systems, following standard TREC-style pooling to reduce judgement bias towards any single retrieval method.

Table 1. BiIR-2026 bilingual corpus and evaluation set statistics

Statistic	Value
Source composition	Samanantar subset / PIB releases / Wikipedia
Document pairs (English-Hindi)	1,860
Total source documents	3,720
PIB press releases	42%
Wikipedia article pairs	31%
Bilingual news-desk reports	27%
Median document length	1,120 words (range 180 to 7,650)
Indexed passages (after chunking)	14,275
Chunk size / overlap	900 characters / 120 characters
Evaluation queries	742
Factoid queries	289 (39%)
Keyword queries	210 (28%)
Conceptual / paraphrastic queries	243 (33%)
Query language direction: English to Hindi (En → Hi)	401 (54%)
Query language direction: Hindi to English (Hi → En)	341 (46%)
Relevance judgement scale	Graded, 0 to 3
Train / validation / test split (queries)	480 (65%) / 112 (15%) / 150 (20%)

Data preprocessing

Documents in DOCX, PDF and HTML formats were converted to plain text (UTF-8) while preserving paragraph boundaries. Hindi text was then normalised using Unicode NFC, and zero-width joiners and control characters were removed to ensure script consistency.

For lexical retrieval, language-specific preprocessing was performed using the Indic NLP Library (Hindi) and spaCy (English), comprising tokenisation, stop-word removal and stemming. The semantic retrieval pipeline instead used normalised raw text without stop-word removal, in order to preserve the contextual information required by transformer-based embeddings. Documents were divided into overlapping passages of 900 characters with a 120-character overlap to maintain semantic continuity. Finally, duplicate passages with a cosine similarity above 0.97 were removed to eliminate redundancy and improve retrieval efficiency.

Query sets and relevance labels

Queries were generated through a two-stage process. An LLM (Llama-3.1-8B-Instruct) first proposed candidate question-answer-source triplets conditioned on each passage, and these candidates were then manually reviewed in English and Hindi. Each query was additionally labelled with a query type (keyword, factoid or conceptual) and a language direction (query language versus relevant-document language) to enable stratified analysis. Inter-annotator agreement was measured on a 15% doubly annotated sample using Cohen's weighted kappa.

System architecture

The proposed system follows a RAG framework comprising two stages: (i) an offline document ingestion pipeline that generates a searchable vector index, and (ii) an online query-response pipeline that combines hybrid retrieval, URF fusion, LLM-based reranking and grounded answer generation. The backend was implemented as a stateless retrieval API with three independently scalable services. A document processor handles text extraction and chunking; an embedding and indexing service generates multilingual sentence embeddings and stores them in a vector database; and a retrieval and ranking service performs hybrid search, score fusion and LLM-based reranking. These components are decoupled through a message queue, which allows each service to scale independently and allows ingestion throughput to be scaled separately from query throughput. The backend is exposed to end users through a purpose-built single-page web application that supports document upload and provides a language-selection control for the desired response language.

Integrated URF-Based Bilingual IR Workflow

When a user submits a query q, the system first determines whether translation is required. If the language of q differs from that of the target documents, q is translated into a pivot language (English) for downstream processing. The query is then processed simultaneously by (i) a semantic retriever, which converts the query into an embedding vector and performs nearest-neighbour search over the vector database, and (ii) a lexical retriever, which tokenises and stems the query and ranks documents using a BM25 index. The top 20 results from each retriever are combined by the URF method to produce a single ranked list, from which the top 10 results are forwarded to an LLM-based reranker that re-evaluates each passage against the

complete query and produces a more accurate relevance ordering. Finally, the highest-ranked passages are grouped by source document and summarised to provide contextual input for the generator LLM. The model generates the final answer from this context, and the answer is then translated back into the user's preferred language. The complete workflow is shown in Figure 1.

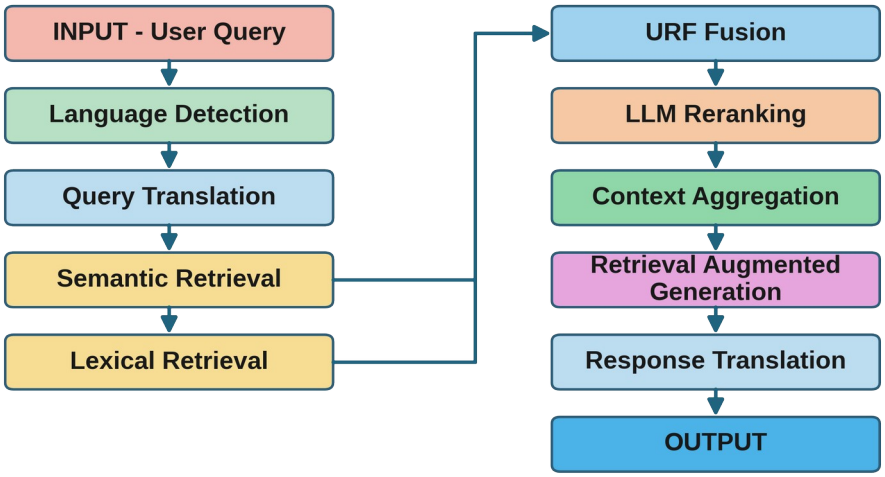

Figure 1. URF-based bilingual IR workflow

Unique Ranking Framework (URF): formulation and hybrid methods

Let q denote a query and d a candidate passage. The semantic similarity is computed as the cosine similarity between the embeddings of q and d produced by the multilingual encoder E:

$$S_{\text{sem}}(q,d) = \frac{\mathbf{E}(q) \cdot \mathbf{E}(d)}{\|\mathbf{E}(q)\| \, \|\mathbf{E}(d)\|} \tag{1}$$

where $S_{sem}(q, d)$ is the semantic similarity score between query q and document (or passage) d; $E(q)$ and $E(d)$ denote the embedding vectors of the query and the document respectively; $E(q) \cdot E(d)$ is the dot product of the two embedding vectors; and $\|E(q)\|$ and $\|E(d)\|$ denote their Euclidean (L2) norms. $S_{sem}(q, d)$ is therefore the cosine similarity between the query and document embeddings, with values ranging from -1 to 1, where higher values indicate greater semantic similarity.

The lexical score is calculated using the Okapi BM25 algorithm, which measures the similarity between q and d over a tokenised, stemmed and stop-word-filtered index. To ensure comparability, BM25 values are normalised to the interval [0, 1] using min-max normalisation within the retrieved candidate pool:

$$S_{\text{lex}}(q,d) = \frac{\text{BM25}(q,d) - \min\limits_{d_j \in D} \text{BM25}(q,d_j)}{\max\limits_{d_j \in D} \text{BM25}(q,d_j) - \min\limits_{d_j \in D} \text{BM25}(q,d_j)} \tag{2}$$

where $S_{lex}(q, d)$ is the normalised lexical relevance score of document (or passage) d for query q; BM25(q, d) is the raw Okapi BM25 score; D denotes the set of retrieved candidate passages; and the min and max operators return the minimum and maximum BM25 scores over all candidates in D, thereby scaling the lexical scores to the interval [0, 1].

Stage 1 (coarse fusion) combines the two normalised signals through a convex mixing weight α in [0, 1]:

$$S_{\text{fuse}}(q,d) = \alpha S_{\text{sem}}(q,d) + (1-\alpha) S_{\text{lex}}(q,d) \tag{3}$$

where $S_{fuse}(q, d)$ is the fused relevance score for document (or passage) d with respect to query q; $S_{sem}(q, d)$ is the normalised semantic similarity score; $S_{lex}(q, d)$ is the normalised lexical relevance score; and $\alpha \in [0, 1]$ is the weighting parameter that controls the relative contribution of semantic and lexical retrieval. Higher values of α place greater emphasis on semantic similarity, whereas lower values increase the influence of lexical matching.

Unlike RRF, which discards raw scores and combines only rank positions, stage-1 fusion in URF operates directly on the normalised similarity scores. This is important when the top-k list is passed to a single LLM reranking stage. The top k candidates (k = 10) are retained and rescored by an LLM cross-encoder, which yields a relevance probability:

$$S_{\text{LLM}}(q,d) = P_{\text{LLM}}(\text{relevant} \mid q,d), \qquad S_{\text{LLM}}(q,d) \in [0,1] \tag{4}$$

where $S_{LLM}(q, d)$ is the relevance score assigned by the LLM to document d for query q, and P_{LLM} is the probability estimated by the LLM.

The final URF score is a second convex combination, governed by the weight β, applied over the LLM-reranked candidate set:

$$\text{URF}(q,d) = \beta S_{\text{fuse}}(q,d) + (1-\beta) S_{\text{LLM}}(q,d) \tag{5}$$

where URF(q, d) is the final unified rank fusion score for query q and document d; $S_{fuse}(q, d)$ is the fused retrieval score obtained by combining semantic and lexical relevance; $S_{LLM}(q, d)$ is the LLM-based relevance score; and $\beta \in [0, 1]$ is the fusion weight.

Both α and β are corpus hyperparameters tuned on the validation split; the sensitivity of end-to-end ranking quality to α is examined below. Documents are finally ranked in descending order of URF(q, d) and grouped by source document before being passed to the generator LLM.

Learning-to-Rank Framework and Training

The mixing weights α and β, together with an exact-entity-match count and a passage length ratio, are learned rather than fixed. Learning was framed as pairwise learning to rank: for each training query q and each pair of candidate passages (d_i, d_j) with graded relevance labels rel_i and

rel_j, a logistic pairwise loss is minimised over a linear scoring function of S_{sem}, S_{lex}, S_{LLM} and the entity-match feature:

$$L = \sum_{(q, d_i, d_j)} \log\left(1 + \exp\left(-\text{sign}(rel_i - rel_j) \cdot (f(d_i) - f(d_j))\right)\right) \quad (6)$$

where f(d) is the weighted linear combination of the four features. The learned coefficients are subsequently renormalised to the convex weights α and β used in Equations 3 and 5. Training used mini-batch gradient descent (Adam, learning rate 1×10^{-3}) over the 480-query (65%) training split, with early stopping on validation NDCG@10 (112 queries, 15%); the held-out 150-query (20%) test split was used only for the final reported metrics.

Implementation Details

The proposed URF framework was implemented in Python 3.11 using the PyTorch deep learning framework. Multilingual sentence embeddings were generated with the Language-agnostic BERT Sentence Embedding (LaBSE) model through the SentenceTransformers library. Dense-vector indexing and approximate nearest-neighbour retrieval were performed using Facebook AI Similarity Search (FAISS) (Johnson et al., 2019) with a Hierarchical Navigable Small World (HNSW) index (Malkov & Yashunin, 2020) for efficient similarity search over the corpus. Lexical retrieval was implemented using the BM25Okapi algorithm from the rank_bm25 package (Brown, 2020). English documents were preprocessed using spaCy (en_core_web_lg) (Honnibal & Montani, 2020), while Hindi normalisation and tokenisation were performed with the Indic NLP Library. Unicode normalisation (NFC), punctuation normalisation and removal of zero-width Unicode characters were applied before indexing.

Documents were segmented into passages of 900 characters with a 120-character overlap, and each embedding was stored as a 768-dimensional floating-point vector. The reranking stage used a cross-encoder large language model (Llama-3.1-8B-Instruct) operating on the top 10 candidates returned by the hybrid retrieval stage. Each query-passage pair was evaluated independently, and the model returned a relevance probability normalised to the interval [0, 1]. The same LLM was used for this step with deterministic decoding (temperature = 0.0).

Training of the pairwise ranking model used the Adam optimiser with an initial learning rate of 1×10^{-3}, a batch size of 32, and early stopping on validation NDCG@10 with a patience of ten epochs.

Complexity and Efficiency Analysis

Semantic retrieval over the HNSW index has an expected query complexity of O(log N) in the number of indexed passages N, while BM25 lookup over an inverted index is O(|q| · average postings length), which is effectively sub-linear in N for typical vocabularies. The dominant cost in the URF pipeline is the LLM reranking stage, which requires one forward pass per candidate passage; restricting this stage to the top k = 10 fused candidates bounds its cost at O(k) LLM calls per query, independent of corpus size N. Table 2 reports the measured end-to-

end latency per pipeline stage on the evaluation corpus (14,275 passages) using a single GPU worker for embedding and reranking.

Table 2. Per-query latency measured by pipeline stage

Pipeline stage	Mean latency (ms)	Complexity (N = corpus size, k = rerank depth)
Query embedding	19	O(1)
Semantic (HNSW) search	24	O(log N)
Lexical (BM25) search	10	O(|q|)
Stage-1 URF fusion	3	$O(k_1 + k_2)$
LLM reranking (k = 10)	414	O(k)
Answer generation and translation	544	O(1) (bounded by output length)
Total end-to-end (median)	1,014	-

Results

The retrieval performance of the Unique Ranking Framework (URF) was assessed on the independent test split of the BiIR-2026 benchmark, consisting of 150 queries excluded from model development and hyperparameter optimisation. URF was compared against seven representative baselines covering dictionary-based retrieval, probabilistic lexical retrieval, dense semantic retrieval, sparse neural retrieval and late-interaction neural retrieval. Retrieval effectiveness was assessed using Precision@10, Recall@10, mean average precision (MAP), mean reciprocal rank (MRR) and normalized discounted cumulative gain at rank 10 (NDCG@10), which together measure retrieval accuracy, ranking quality and early precision.

To ensure a fair comparison, identical query sets, document collections, preprocessing pipelines and evaluation scripts were used for all retrieval methods. Hyperparameters of the proposed framework were optimised exclusively on the validation split, while the test partition remained completely isolated until the final evaluation.

URF Hybrid attained the highest performance across all five evaluation metrics (Table 3; Figure 2). Compared with the dictionary-based CLIR baseline, URF improved MRR by 83.3% and NDCG@10 by 105.0%. Compared with Indic-ColBERT, the strongest baseline, URF improved MRR by 8.5%, NDCG@10 by 7.9% and MAP by 7.0%. These improvements were statistically significant ($p < 0.01$) across the test split.

Table 3. Overall retrieval performance of URF and the baseline systems on the BiIR-2026 test split

Method	Precision@10	Recall@10	MAP	MRR	NDCG@10
Dictionary CLIR	0.41	0.48	0.36	0.42	0.40
BM25 (Okapi)	0.45	0.52	0.42	0.46	0.48
LaBSE (pure embedding)	0.56	0.61	0.51	0.51	0.55
DPR (dense passage)	0.60	0.65	0.58	0.57	0.64
SPLADE (sparse neural)	0.61	0.65	0.64	0.61	0.66
ColBERT (late interaction)	0.65	0.71	0.68	0.65	0.70
Indic-ColBERT	0.67	0.75	0.71	0.71	0.76
URF Hybrid (α = 0.6)	0.70	0.78	0.76	0.77	0.82

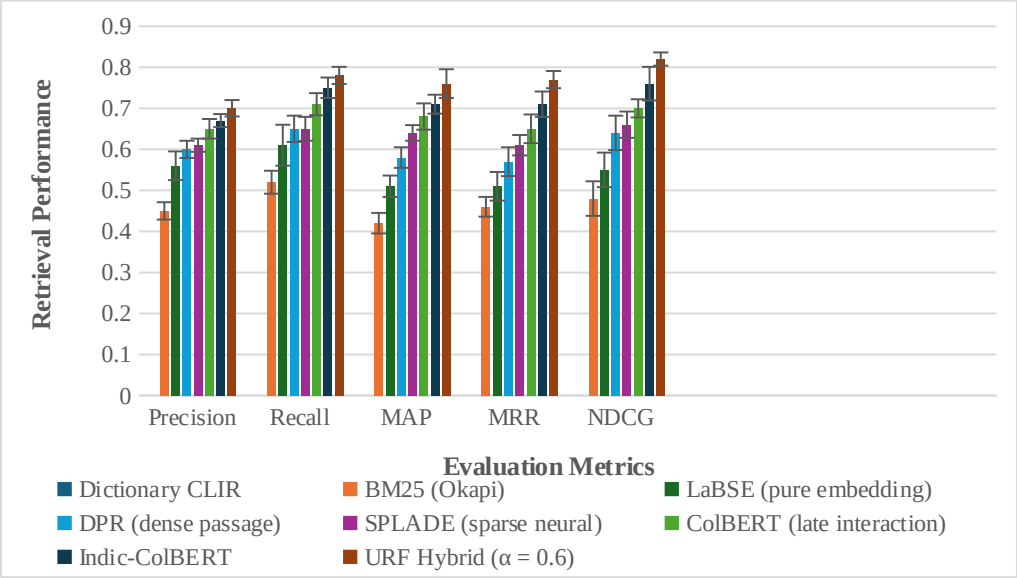

Figure 2. Overall retrieval performance comparison across methods

Statistical Evaluation

To determine whether the improvements obtained by URF were statistically significant rather than the product of random variation, statistical testing was performed using paired comparisons of query-level retrieval scores. Because every retrieval method was evaluated on the identical query set, a paired statistical test was appropriate for comparing system performance. Statistical significance was assessed using the Wilcoxon signed-rank test at $p < 0.05$. In addition, 95% confidence intervals were estimated by bootstrap resampling with 10,000 iterations to establish the robustness of the reported metrics.

The largest single improvement occurred at the transition from conventional dictionary-based retrieval to the first embedding-based model (LaBSE), highlighting the advantage of translation-free semantic matching for CLIR. Subsequent neural retrieval models such as DPR, SPLADE, ColBERT and Indic-ColBERT delivered smaller incremental gains. The URF hybrid

framework, which combines lexical-semantic fusion with LLM-based reranking, produced a further improvement over this progression.

Analysis by Query Type and Language Direction

Table 4 reports MRR and NDCG@10 for the two strongest baselines and the two URF configurations across the three query types (factoid, keyword and conceptual) and the two language directions (English query to Hindi document, and Hindi query to English document).

Table 4. MRR and NDCG@10 by query type and query-answer language direction

Method	Factoid (MRR / NDCG)	Keyword (MRR / NDCG)	Conceptual (MRR / NDCG)	En → Hi (MRR / NDCG)	Hi → En (MRR / NDCG)
ColBERT	0.75 / 0.78	0.72 / 0.74	0.60 / 0.64	0.70 / 0.73	0.68 / 0.71
Indic-ColBERT	0.78 / 0.81	0.75 / 0.77	0.63 / 0.66	0.73 / 0.75	0.71 / 0.73
URF (fusion only, no LLM rerank)	0.79 / 0.82	0.77 / 0.79	0.68 / 0.71	0.76 / 0.78	0.74 / 0.76
URF Hybrid (full, α = 0.6)	0.82 / 0.84	0.80 / 0.82	0.74 / 0.77	0.80 / 0.82	0.77 / 0.79

The confidence intervals associated with the retrieval metrics were narrow across all evaluations, indicating stable retrieval performance for the proposed framework. For the full URF pipeline, the estimated 95% confidence interval for MRR was 0.78 ± 0.02 and that for NDCG@10 was 0.80 ± 0.02, indicating low variability across evaluation queries.

The largest gain was observed for conceptual, paraphrastic queries (+0.06 MRR and +0.06 NDCG@10 relative to the fusion-only configuration), which supports the hypothesis that LLM reranking contributes most where lexical overlap between query and passage is low and semantic reasoning is required to establish relevance. Both the fusion-only configuration and the complete URF pipeline performed consistently better on En → Hi retrieval than on Hi → En retrieval. This asymmetry may be due to the English-dominant pretraining of the multilingual embedding model, which produces more discriminative and better calibrated semantic representations for English queries than for Hindi queries.

Ablation and Sensitivity Studies

The sensitivity of end-to-end NDCG@10 to the stage-1 fusion weight α (Equation 3) was examined by holding β fixed at its tuned value ($\beta = 0.55$) and sweeping α from 0 to 1 (pure lexical to pure semantic) on the validation split. The results are shown in Figure 3.

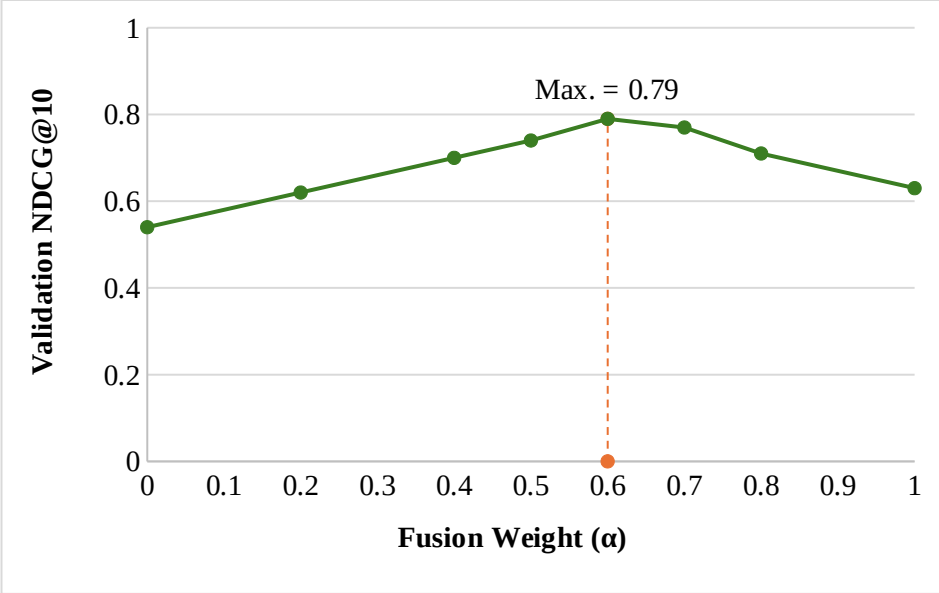

Figure 3. Sensitivity of validation NDCG@10 to the fusion weight α (β fixed at 0.55)

NDCG@10 peaks at α = 0.6, and performance is robust to moderate mistuning of α: within ± 0.1 of the optimum, NDCG@10 degrades by less than 0.02. Performance degrades more sharply towards pure lexical search (α = 0) than towards pure semantic search (α = 1), which is consistent with a cross-lingual setting in which lexical overlap is a weak signal for En-Hi passage pairs.

The influence of the LLM reranking stage on retrieval effectiveness and computational efficiency was examined by varying the reranking depth k, that is, the number of top-ranked candidates from stage-1 fusion forwarded to the LLM for reranking (Figure 4). Reranking depths beyond k = 10 provided only marginal improvements in NDCG@10, whereas end-to-end retrieval latency increased linearly with k. A value of k = 10 therefore offers the most suitable balance between retrieval effectiveness and computational cost, and it was adopted as the default reranking depth in the proposed URF framework and used throughout the experiments.

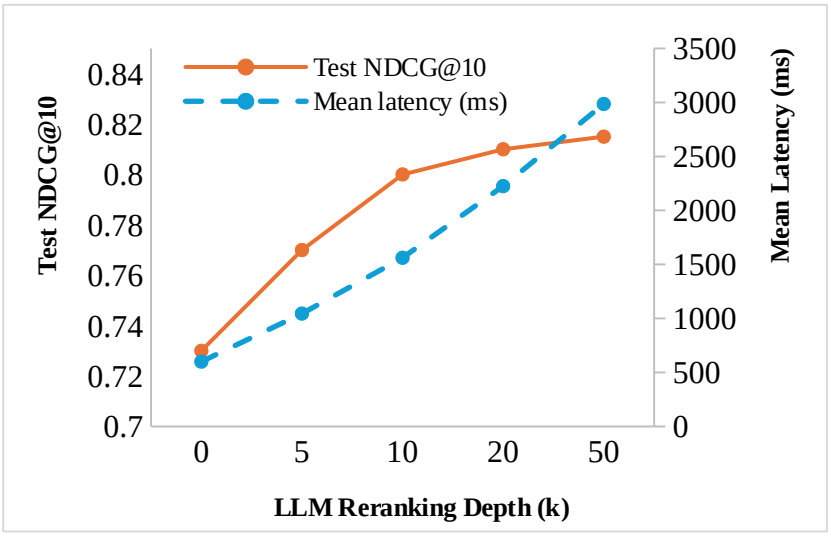

Figure 4. Effect of LLM reranking depth k on NDCG@10 and mean latency

Despite the improvements attained by URF, a number of retrieval failure cases were identified during qualitative analysis. Most retrieval errors occurred for highly ambiguous entity names, transliterated proper nouns, numerically dense government notifications, domain-specific abbreviations and short keyword queries.

Dense retrieval occasionally returned semantically related but factually incorrect passages because of contextual similarity, whereas lexical retrieval failed whenever English and Hindi documents shared minimal surface vocabulary. Although LLM-based reranking substantially reduced these errors, failure cases persisted whenever the retrieved candidate passages themselves lacked sufficient supporting evidence. These observations indicate that future improvements should focus primarily on candidate generation rather than on reranking alone.

Discussion

The experimental evaluation shows that the improvement in bilingual retrieval quality arises from the combined contributions of semantic retrieval, lexical retrieval and contextual reranking rather than from any single component alone. The transition from dictionary-based retrieval to multilingual dense embeddings delivered the largest single improvement in retrieval effectiveness, confirming that semantic representation learning substantially reduces the lexical mismatch between English and Hindi. Dense retrieval alone, however, was insufficient for exact entity matching and numerical expressions, which motivates the integration of lexical retrieval within the proposed hybrid framework.

Unlike reciprocal rank fusion, which combines only document ranks (Bruch et al., 2023; Clipa & Di Nunzio, 2020), URF integrates normalised semantic and lexical similarity scores directly before contextual reranking by a large language model. This score-level fusion preserves richer relevance information while allowing the contribution of each retrieval component to be optimised through corpus-specific weighting parameters. Consequently, URF consistently outperformed all baseline retrieval systems across multiple evaluation metrics.

The second source of gain was the combination of lexical and semantic signals in the URF fusion stage together with LLM reranking, which recovers the complementary strength of exact term and entity matching that pure dense retrieval can miss, while adding semantic reasoning capacity. The per-query-type breakdown showed that this gain is concentrated in conceptual queries, where paraphrase and implicit relevance are more common, rather than in factoid queries, where strong entity overlap already places dense and late-interaction models close to their ceiling.

Stage-1 fusion in URF generalises RRF by operating on normalised raw scores rather than on rank positions alone. Figure 3 shows that the resulting fusion weight has a broad optimum rather than requiring precise tuning, which is a practical advantage over rank-only fusion methods whose behaviour is difficult to interpret in terms of a single mixing coefficient. Rather than relying solely on a stronger embedding model, as many multilingual RAG systems do, URF improved retrieval performance by integrating hybrid lexical-semantic retrieval and LLM-based reranking over the same multilingual encoder. The observed gains suggest that optimising the retrieval pipeline can be as beneficial as upgrading the embedding model itself.

Improved fusion and reranking mechanisms and stronger embeddings are complementary, each improving retrieval quality in a different way, which indicates that their combination offers considerable potential for further improvement.

Limiting LLM reranking to a small set of high-confidence candidates is essential to maintaining the feasibility of URF for interactive retrieval. The results show that reranking and response generation together account for most of the end-to-end processing time, while increasing the reranking depth beyond k = 10 yields only marginal improvements in NDCG@10 despite a steady increase in latency. This implies that k should be treated as a configurable trade-off between retrieval quality and response time rather than fixed at the largest feasible value. The optimal value of k may also depend on query type: factoid queries, for which the improvement is relatively modest, could be processed with a smaller k to reduce latency, whereas conceptual queries benefited substantially from LLM-based reranking (Lu et al., 2025) and are better suited to a deeper reranking stage. This highlights the potential of query-adaptive reranking strategies to preserve retrieval quality without incurring unnecessary computational cost.

From a deployment perspective, the proposed framework is computationally feasible because expensive LLM inference is restricted to the highest-ranked candidate passages. The hybrid retrieval stage effectively reduces the search space before contextual reranking, allowing interactive response times while maintaining high accuracy. This architecture is therefore well suited to government portals, multilingual digital libraries, knowledge portals, legal repositories and public-service question-answering systems, where both retrieval quality and response latency are critical.

Conclusion and Future Work

This study proposed a Unique Ranking Framework (URF) for bilingual English-Hindi cross-lingual information retrieval that integrates lexical retrieval, semantic retrieval and LLM-based reranking within a RAG architecture. Unlike traditional reciprocal rank fusion methods, which combine only rank positions, URF performs two-stage score-level fusion in which normalised semantic and lexical similarity scores are merged using an optimised convex weighting strategy, and the fused candidates are then reranked by an LLM. To evaluate the framework, a new graded-relevance benchmark, BiIR-2026, consisting of 1,860 aligned English-Hindi document pairs and 742 queries, was developed from the Samanantar corpus, PIB press releases and comparable Wikipedia articles. URF was compared with seven representative retrieval baselines on Precision@10, Recall@10, MAP, MRR and NDCG@10, and extensive ablation studies examined the impact of the fusion weights, reranking depth, latency and retrieval effectiveness. The framework delivers a significant improvement in retrieval effectiveness and also provides a practical architecture for multilingual retrieval systems targeting low-resource Indic languages.

Future work will concentrate on improving candidate generation rather than reranking alone, on query-adaptive selection of the reranking depth k, and on extending both the framework and the BiIR-2026 benchmark to additional Indic languages.

Data Availability

The BiIR-2026 benchmark developed in this study will be made publicly available upon publication through a public repository. Source documents drawn from publicly available datasets remain subject to their original licences.

Conflict of Interest

The authors declare that they have no competing interests related to this work.

References

Arslan, M., Ghanem, H., Munawar, S., & Cruz, C. (2024). A survey on RAG with LLMs. *Procedia Computer Science, 246*, 3781-3790.

Bhattacharya, P., Goyal, P., & Sarkar, S. (2018). Using communities of words derived from multilingual word vectors for cross-language information retrieval in Indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing, 18*(1), 1-27.

Brown, D. (2020). *Rank-BM25: A collection of BM25 algorithms in Python* [Computer software].

Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems, 42*(1), 1-35.

Chavhan, S., & Dharmik, R. (2025). Optimizing learning to rank models with neural network-based feature selection techniques. *Journal of Information and Optimization Sciences, 46*(2), 541-551.

Clipa, T., & Di Nunzio, G. M. (2020). A study on ranking fusion approaches for the retrieval of medical publications. *Information, 11*(2), 103.

Dave, B., & Majumder, P. (2026). SqCLIRIL: Spoken query cross-lingual information retrieval in Indian languages. *Pattern Recognition Letters, 199*, 288-294.

Dwivedi, S. K., & Chandra, G. (2016). A survey on cross-language information retrieval. *International Journal on Cybernetics & Informatics, 5*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Gao, L., Dai, Z., Chen, T., Fan, Z., Van Durme, B., & Callan, J. (2021). Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*. Springer.

Goworek, R., Macmillan-Scott, O., & Özyiğit, E. B. (2025). Bridging language gaps: Advances in cross-lingual information retrieval with multilingual LLMs. *arXiv preprint arXiv:2510.00908*.

Hambarde, K. A., & Proenca, H. (2023). Information retrieval: Recent advances and beyond. *IEEE Access, 11*, 76581-76604.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., & Järvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval, 7*(1), 99-119.

Honnibal, M., & Montani, I. (2020). *spaCy: Industrial-strength natural language processing in Python* [Computer software].

Hsu, H.-L., & Tzeng, J. (2025). DAT: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation. *arXiv preprint arXiv:2503.23013*.

Iswarya, P., & Radha, V. (2012). Cross language text retrieval: A review. *International Journal of Engineering Research and Applications*, *2*, 1036-1043.

Jagarlamudi, J., & Kumaran, A. (2007). Cross-lingual information retrieval system for Indian languages. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535-547.

Jones, K. S., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Information Processing & Management*, *36*(6), 809-840.

Kafi, A., Saha, S., & Shahriar, N. (2026). Weighted reciprocal rank fusion RAG for context-aware DoS attack mitigation. In *2026 IEEE 23rd Consumer Communications & Networking Conference (CCNC)*. IEEE.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Katta, E., & Arora, A. (2015). An improved approach to English-Hindi based cross language information retrieval system. In *2015 Eighth International Conference on Contemporary Computing (IC3)*. IEEE.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management*, *41*(3), 523-547.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Lu, T., Zhou, Z., Wang, J., & Wang, Y. (2025). A large language model-based approach for personalized search results re-ranking in professional domains. *Academia Nexus Journal*, *4*(1).

Luan, Y., Eisenstein, J., Toutanova, K., & Collins, M. (2021). Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, *9*, 329-345.

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824-836.

Nair, A. R., & Gupta, D. (2026). Direct preference optimization in LLMs for Indian language translation guided by COMET quality estimation. *Procedia Computer Science*, *283*, 189-201.

Nawaz, A. (2025). Advancements in natural language processing for multilingual information retrieval systems. *Multidisciplinary Research in Computing Information Systems*, *5*(3), 197-216.

Nic Gearailt, D. (2005). *Dictionary characteristics in cross-language information retrieval*.

Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *arXiv preprint arXiv:2507.18910*.

Puranegedara, I., Chathumina, T., Ranathunga, N., De Silva, N., Ranathunga, S., & Thayaparan, M. (2025). Utilizing multilingual encoders to improve large language models for low-resource languages. In *Moratuwa Engineering Research Conference (MERCon)*. IEEE.

Rahimi, R., Shakery, A., & King, I. (2015). Multilingual information retrieval in the language modeling framework. *Information Retrieval Journal*, *18*(3), 246-281.

Rajaei, D., Taheri, Z., & Fani, H. (2024). Enhancing RAG's retrieval via query backtranslations. In *International Conference on Web Information Systems Engineering*. Springer.

Rao, A., Alipour, H., & Pendar, N. (2025). Rethinking hybrid retrieval: When small embeddings and LLM re-ranking beat bigger models. *arXiv preprint arXiv:2506.00049*.

Robertson, S., & Zaragoza, H. (2009). *The probabilistic relevance framework: BM25 and beyond* (Vol. 4). Now Publishers.

Sciavolino, C., Zhong, Z., Lee, J., & Chen, D. (2021). Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Seetha, A., Das, S., & Kumar, M. (2007). Evaluation of the English-Hindi cross language information retrieval system based on dictionary based query translation method. In *10th International Conference on Information Technology (ICIT 2007)*. IEEE.

Shahzad, K., Ashraf, A., & Salama, M. (2026). Using AI for Islamic jurisprudence questions: Comparative analysis of domain-specific retrieval-augmented generation (RAG) vs. generic LLM. *Journal of Integrated Sciences*, *6*(3), 1-51.

Sharma, V. K., & Mittal, N. (2016). Cross lingual information retrieval (CLIR): Review of tools, challenges and translation approaches. In *Information Systems Design and Intelligent Applications: Proceedings of the Third International Conference INDIA 2016* (Vol. 1). Springer.

Sree, K. D., Gupta, D., & Nair, A. R. (2026). Domain-specific retrieval-augmented generation for English-to-Hindi machine translation using BLOOMZ-3B. *Procedia Computer Science*, *283*, 202-212.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., & Ren, Z. (2023). Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wang, S., Zhuang, S., & Zuccon, G. (2024). Large language models based stemming for information retrieval: Promises, pitfalls and failures. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Zhang, W., & Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, *13*(5), 856.

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, *45*(1), 1-44.

Zhu, X., & Klabjan, D. (2020). Listwise learning to rank by exploring unique ratings. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., & Wen, J.-R. (2025). Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, *44*(1), 1-54.