

A Comparative Study of Lexical, Dense, and Hybrid Retrieval for Bilingual Graduation Proposal Similarity Detection

Marwa Eddakhli¹, Shaima Salem Rashed Embashi¹, and Tairy Mousa Ali¹

Programming Department, College of Computer Technology Tripoli, Libya
{marwa.eddakhli, shaima22102004, tairymousa}@gmail.com

Abstract. We compare three retrieval strategies for bilingual (Arabic–English) graduation project proposal similarity detection: a lexical baseline (BM25), a dense pipeline (fine-tuned Multilingual-E5 Bi-Encoder with Cross-Encoder re-ranking), and a hybrid pipeline (Reciprocal Rank Fusion). Models were evaluated on a strict cross-lingual gold standard of 300 real Arabic proposal pairs translated to English. BM25 exhibited severe performance degradation, approximating random-rank performance due to vocabulary mismatch. The Dense pipeline outperformed the Hybrid pipeline under standard fusion parameters. While highly tuned hybrid configurations achieved statistical parity, applying the principle of parsimony, we conclude that within the scope of this proof-of-concept study ($N = 300$, single institution, synthetic fine-tuning data), the pure Dense pipeline is the most operationally viable approach for this task, pending validation at institutional scale. It achieves equivalent peak accuracy without the architectural overhead of maintaining a lexical index. Qualitative evaluation revealed that hybrid fusion introduces spurious BM25 keyword matches, producing false positives on novel queries, whereas the pure dense pipeline correctly rejects them.

Keywords: Bilingual Retrieval · Semantic Similarity · Cross-Encoder · Multilingual-E5

1 Introduction

Academic institutions that offer computer science programmes accumulate hundreds of graduation project proposals each year. Supervisors review these proposals manually—a labour-intensive process susceptible to oversight errors under high submission volumes, exhibiting poor scalability. This risk is operationally significant: students may independently submit proposals that are semantically equivalent separated only by surface-level rephrasing, motivating automated text similarity approaches [1,2]. In bilingual environments such as Libyan universities, where proposals arrive in both Arabic and English, this challenge is further compounded. A keyword search cannot detect that an Arabic proposal on “smart irrigation using IoT” describes the same system as an English proposal on “automated farm watering with sensors.” Despite sharing no surface tokens, the two proposals describe semantically equivalent engineering objectives.

Three families of retrieval techniques compete for this task. Lexical methods, led by BM25 [3], rank documents by weighted term overlap; they offer computational efficiency and interpretability but lack the capacity to capture paraphrase and synonym relationships, and when the query and document share no surface tokens, the BM25 score is identically zero—a deterministic failure mode in cross-lingual settings. Dense methods encode text into a shared vector space via multilingual transformers [4,5,6], then retrieve by maximum inner-product search on L_2 -normalised vectors; they handle paraphrase and cross-lingual matching but require GPU inference and careful fine-tuning. Hybrid methods fuse both signals—typically through Reciprocal Rank Fusion (RRF) [7]—on the premise that the combination exceeds either component.

The optimal strategy should be determined empirically, independent of architectural sophistication. We designed an experiment to resolve this question through a controlled proof-of-concept. We constructed 6,000 synthetic bilingual proposals (3,000 Arabic, 3,000 English) to fine-tune a Multilingual-E5 Bi-Encoder [6]. We then collected 300 real Arabic graduation proposals from the College of Computer Technology, Tripoli, translated them to English via DeepL with human review, and used the resulting 300 aligned pairs as a strict held-out gold standard. Three pipelines were evaluated: (1) a pure Lexical baseline (BM25), (2) a Dense pipeline (fine-tuned Multilingual-E5 + Cross-Encoder re-ranker), and (3) a Hybrid pipeline (BM25 + Dense + RRF + Cross-Encoder).

We evaluate these three pipelines against this gold standard to determine: (a) whether BM25 retains any viability in the cross-lingual setting, (b) whether hybrid fusion adds measurable value over pure dense retrieval when lexical signals are absent, and (c) whether any performance gain from hybrid architectures justifies the additional complexity of maintaining a separate lexical index and tuning fusion hyperparameters.

This study makes three contributions:

1. **Resource Construction:** The creation of a rigorously translated and domain-expert-verified gold standard of 300 cross-lingual Arabic–English proposal pairs, alongside a 6,000-pair synthetic fine-tuning corpus.
2. **Empirical Finding:** A controlled demonstration that in strict zero-vocabulary-overlap cross-lingual settings, standard Hybrid retrieval degrades performance compared to pure Dense retrieval due to BM25 noise injection, with the degradation confirmed by statistical significance testing.
3. **Methodological Insight:** A qualitative and quantitative validation showing that pure Dense pipelines offer superior operational robustness by correctly rejecting false positives on novel queries that lexical fusion inadvertently promotes.

2 Background

This section introduces the four retrieval components evaluated in the study: the BM25 lexical scorer, multilingual transformer encoders, the Bi-Encoder and Cross-Encoder architectures, and Reciprocal Rank Fusion.

2.1 BM25

BM25 is a bag-of-words ranking function derived from the Probabilistic Relevance Framework [3]. Given a query q and document d , BM25 computes a relevance score by weighting each shared term’s frequency in d against its inverse document frequency across the corpus, with a saturation parameter k_1 and a length normalisation factor b . The function is efficient—scoring requires only an inverted index lookup—and it remains a strong baseline in monolingual benchmarks. Its failure mode is deterministic: when q and d share no tokens, the score is exactly zero. In a cross-lingual setting where Arabic queries search an English corpus (or vice versa), this failure is the default rather than the exception.

2.2 Multilingual Transformer Models

BERT [4] introduced deep bidirectional pre-training on large text corpora via masked language modelling, producing contextual representations that transfer well to downstream tasks. While single-language models like AraBERT [8] and CAMEL-BERT [9] achieve state-of-the-art performance on monolingual Arabic NLP tasks, they cannot natively encode English queries into a shared space. Multilingual extensions—mBERT (trained on 104 languages) and XLM-RoBERTa [5] (trained on 2.5 TB of CommonCrawl data across 100 languages)—solve this by aligning semantically equivalent sentences from different languages in a shared space. This cross-lingual alignment emerges from the overlap in subword vocabularies and the shared transformer parameters, not from parallel corpora or explicit translation signals. The property is central to our Dense pipeline: it allows a single encoder to map both Arabic and English proposals into the same vector space.

2.3 Bi-Encoders and Dense Retrieval

Sentence-BERT [10] adapts the transformer architecture into a Bi-Encoder: two input texts are encoded independently into fixed-dimensional dense vectors, and their similarity is measured by cosine distance. Because corpus documents can be encoded offline and stored in a vector index such as FAISS, retrieval reduces to calculating inner products. In our experimental setup with 300 documents, we utilise `IndexFlatIP` to perform exact exhaustive inner-product search. This index computes $\mathcal{O}(N \cdot d)$ dot products per query (where $d = 768$), which is trivially fast only due to the negligible corpus scale. Scaling to an anticipated institutional collection of 5,000+ proposals would necessitate an Approximate Nearest Neighbour index such as HNSW—for instance, using typical default parameters ($M = 16$, $ef_construction = 200$) as documented in the FAISS library [11]—for sub-linear retrieval. The Multilingual-E5 model [6] extends the Bi-Encoder approach through weakly-supervised contrastive pre-training on multilingual text pairs drawn from web data. E5 produces 768-dimensional embeddings and achieves competitive cross-lingual alignment on standard benchmarks without requiring parallel sentence supervision. We adopt `intfloat/multilingual-e5-base` as our Bi-Encoder backbone throughout this work.

2.4 Cross-Encoder Re-Ranking

A Bi-Encoder compresses each text into a single vector, discarding fine-grained token interactions. Cross-Encoders [10] recover this information by concatenating the query and document into a single input sequence, processing both jointly through a transformer with full cross-attention. The output is a scalar relevance score that captures token-level alignment between the two texts. This yields higher ranking accuracy at the cost of quadratic complexity in sequence length, making Cross-Encoders impractical as first-stage retrievers but highly effective as re-rankers applied to a short candidate list. In our pipeline, we employ the multilingual `cross-encoder/mmarco-mMiniLMv2-L12-H384-v1` (trained on the mMARCO dataset) as our re-ranker.

2.5 Reciprocal Rank Fusion

Reciprocal Rank Fusion (RRF) [7] merges multiple ranked lists into a single ordering without requiring score calibration. For a document at rank r in list i , RRF assigns a contribution of $1/(k + r)$, where k is a constant that dampens the influence of high-ranked outliers. The fused score for each document is the sum of its contributions across all input lists. RRF is parameter-light (only k must be set) and has been shown to outperform both Condorcet voting and individual rank-learning methods on standard TREC collections [7]. We use RRF to combine the BM25 and Bi-Encoder candidate lists in the Hybrid pipeline.

3 Related Work

Research on bilingual retrieval and academic text similarity draws from three overlapping threads: hybrid sparse–dense architectures, cross-lingual dense embeddings, and similarity detection in scholarly documents.

Hybrid Sparse–Dense Retrieval. Lexical retrievers and dense encoders typically exhibit complementary strengths. Modern sparse models like SPLADE [12,13] and dense architectures like ColBERTv2 [14] have further narrowed the gap, but fusion often remains the robust default. Recent literature on hybrid sparse–dense architectures catalogues the conditions under which fusion outperforms either component alone. Kuzi et al. and Lin et al. demonstrated that fusion gains are most pronounced when both signals retain baseline competence [15,16]. Lin et al. proposed a dense representation framework that internalises both lexical and semantic matching signals into a single index [16]. These findings motivated the Hybrid pipeline in our study. However, our results demonstrate a critical violation of fusion assumptions: in strict cross-lingual settings lacking shared vocabulary, the lexical component entirely loses competence and introduces spurious candidates that degrade the fusion stage.

Cross-Lingual Dense Embeddings. Mapping text from multiple languages into a shared vector space is the enabling mechanism behind language-agnostic retrieval. Recent state-of-the-art models like mContriever [17] have

demonstrated that unsupervised contrastive learning can yield highly competitive multilingual dense retrievers. This builds on foundational multilingual dense retrieval baselines such as mDPR and Mr. TyDi [18,19], which first demonstrated the efficacy of cross-lingual dense encoders. Huang extended this line of work by modelling cross-lingual knowledge transfer explicitly within multilingual retrieval systems, demonstrating that a single query encoder can retrieve documents in a language it was not specifically trained on [20]. The Multilingual-E5 model we adopt builds on these foundations through large-scale weakly-supervised contrastive pre-training, aligning 100+ languages without parallel supervision [6]. While contemporary generative variants such as E5-mistral-7b [21] set the current state-of-the-art on multilingual benchmarks, the base E5 model remains the computationally tractable choice for institutional embedding deployments. Furthermore, the robust evaluation of such models is increasingly reliant on standardised zero-shot retrieval suites like the BEIR benchmark [22] and the MTEB (Massive Text Embedding Benchmark) leaderboard [23], which contextualise cross-domain generalisation and model selection claims.

Text Similarity in Academic Documents. A smaller body of work targets similarity detection within scholarly corpora specifically. Donner evaluated multiple similarity methods for identifying constituent articles in cumulative dissertation theses, using a short-text bilingual dataset, and found that no single method dominated across all evaluation conditions [24]. Zhang proposed a fragment-merging algorithm for bilingual text similarity that complements neural approaches in structured document types [25]. On the linguistic side, Alinasab et al. analysed lexical cohesion patterns in bilingual and multilingual graduate theses, reporting that writers with broader multilingual competence produce richer inter-textual semantic connections [26]—an observation consistent with the structured overlap we exploit in our proposal corpus. Wang et al. recently demonstrated that pre-trained language models can establish robust frameworks for semantic document comparison, underpinning much of the subsequent research in academic similarity detection [27].

Benchmarking and Evaluation Gaps. Recent benchmarking efforts, notably the MIRACL dataset [28], have stressed the need for standardised multilingual evaluation protocols, particularly for under-resourced language pairs. Arabic–English academic retrieval is one such under-served pair: no prior study provides a controlled comparison of lexical, dense, and hybrid pipelines on a bilingual corpus of graduation proposals with a strict cross-lingual gold standard. Our work fills this gap.

4 Proposed Methodology

We frame bilingual proposal similarity detection as an information retrieval task: given a new query proposal q , retrieve and rank the most semantically related proposals from a historical corpus $\mathcal{C}$ that may contain documents in Arabic, English, or both. The system is organised into four stages—data ingestion, first-

stage retrieval, optional fusion, and Cross-Encoder re-ranking—which are combined in three configurations to isolate the contribution of each component.

4.1 Data Ingestion and Text Representation

Each proposal is decomposed into structured rubrics: Title, Problem Statement, Proposed Solution, Objectives, and Technologies. These fields are concatenated into a single text representation t_i per proposal. Arabic text undergoes rigorous normalisation via standard regular expression Unicode mapping: diacritics are stripped, alef variants (alef with hamza or madda) are unified to bare alef, taa marbuta is mapped to haa, Farsi yeh is disambiguated and mapped to Arabic yeh, non-standard numerals are converted to Western Arabic digits, and whitespace is collapsed. We explicitly avoided aggressive stemming (e.g., Khoja or ISRI stemmers) because graduation proposals rely heavily on exact technical terminology, and stemming risks conflating distinct technological concepts. For example, an Arabic morphological stemmer might reduce both ‘networking’ (*shabakat*) and ‘interlacing/entangling’ (*tashabuk*) to the same linguistic root (*sh-b-k*), thereby destroying the precise technical distinction required for proposal matching. This conservative normalisation reduces surface variation without altering semantics, ensuring consistent tokenisation for both BM25 and the subword tokeniser of the Bi-Encoder.

4.2 First-Stage Retrieval

Two independent retrievers produce candidate lists from the corpus:

Lexical Retriever (BM25). The corpus is tokenised by whitespace and indexed with BM25Okapi. We explicitly chose not to apply an Arabic morphological segmenter (e.g., Faraji or Buckwalter) to maintain a language-agnostic pipeline, although we acknowledge that aggressive segmentation might have marginally improved lexical recall at the cost of breaking multi-word technical terms. For a query q , BM25 returns the top- K documents ranked by term-frequency scoring ($K = 50$). Both Arabic and English tokens are indexed in a single inverted index without language-specific stemming.

Dense Retriever (Multilingual-E5 + FAISS). Each proposal $t_i \in \mathcal{C}$ is encoded into a 768-dimensional vector $\mathbf{v}_i$ using a fine-tuned Multilingual-E5 Bi-Encoder (`intfloat/multilingual-e5-base`). Embeddings are L_2 -normalised and stored in a FAISS `IndexFlatIP` index [11], so that retrieval reduces to a maximum inner-product search. The query is encoded at inference time and the top- K nearest vectors are returned ($K = 50$).

Bi-Encoder Fine-Tuning. The base Multilingual-E5 model is domain-adapted on 6,000 synthetic proposals using *Multiple Negatives Ranking Loss* (MNRL) [29]. The loss is formally defined as:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{\text{sim}(\mathbf{q}_i, \mathbf{d}_i^+)/\tau}}{\sum_{j=1}^B e^{\text{sim}(\mathbf{q}_i, \mathbf{d}_j)/\tau}} \quad (1)$$

where $B = 16$ is the batch size, $\tau = 0.05$ is a temperature scaling parameter consistent with standard contrastive thresholds [30], $\mathbf{q}_i$ and $\mathbf{d}_i^+$ are embeddings of a positive pair, and $\mathbf{d}_j$ represents in-batch negatives. Due to hardware constraints (32 GB RAM and RTX 3060 Ti), our training batch size was restricted to $B = 16$, providing only 15 in-batch negatives per positive pair. While recent contrastive learning literature demonstrates that larger batch sizes (e.g., 128+) significantly improve representations by exposing the model to more diverse negatives [30,6], our restricted batch size nevertheless proved sufficient to induce a highly discriminative embedding space for this specific domain, without requiring computationally expensive hard-negative mining. Positive training pairs consist of two synthetic proposals *independently generated* to describe the same project concept. Specifically, for each cross-lingual pair, the Arabic and English proposals were produced by separate Gemini Pro generation calls that shared a common semantic concept (e.g., “smart irrigation system using IoT sensors for arid environments”) but received no access to each other’s output. The two proposals were therefore *not* direct translations: they differ in phrasing, sentence structure, and the specific technical details elaborated, while preserving the same core engineering objective. This independent generation strategy means the Bi-Encoder must learn genuine cross-lingual semantic alignment rather than surface-level translational correspondence, since the positive pairs share conceptual intent but not lexical or syntactic structure. Negative pairs are entirely derived from in-batch negatives (i.e., other proposals within the same batch of 16). These pairs are constructed within each domain (Programming, Control, Networks) across three language combinations: Arabic–English, Arabic–Arabic, and English–English. This ensures the model learns both cross-lingual and within-language similarity. Fine-tuning runs for a maximum of 10 epochs with a learning rate of 2×10^{-5} utilising a linear schedule with a 10% warmup. We monitored convergence on a 15% held-out validation split, applying early stopping with a patience of 3 epochs based on validation loss; under these conditions, training converged in approximately 4.5 hours of wall-clock time.

4.3 Candidate Fusion (Hybrid Only)

In the Hybrid configuration, the BM25 and Dense candidate lists are merged via Reciprocal Rank Fusion with $k = 40$. For each document d , the fused score is computed as:

$$\text{RRF}(d) = \sum_i \frac{1}{k + r_i} \quad (2)$$

where r_i is the rank of document d in list i . The fused list is truncated to the top 20 candidates before re-ranking.

4.4 Cross-Encoder Re-Ranking

The top 20 candidates (from the Dense retriever or the RRF fusion, depending on the pipeline) are re-ranked by a multilingual Cross-Encoder (**cross-encoder**/

mmarco-mMiniLMv2-L12-H384-v1). The Cross-Encoder concatenates the query and each candidate into a single transformer input and outputs a scalar relevance score. Candidates are re-ordered by this score. The Cross-Encoder is used off-the-shelf, without domain-specific fine-tuning. This asymmetry (a fine-tuned Bi-Encoder paired with a zero-shot Cross-Encoder) was a practical necessity due to the quadratic computational cost of full cross-attention fine-tuning on 6,000 pairs, and is justified by the mMARCO model’s robust pre-existing multilingual alignment. Specifically, during qualitative analysis, we observed that the zero-shot Cross-Encoder score distributions separated positive and negative pairs adequately (mean scores $\approx +10.5$, $SD = 1.8$ for gold matches versus mean ≈ -2.5 , $SD = 3.2$ for irrelevant candidates; score overlap coefficient < 0.05), empirically supporting its applicability to the proposal domain. Figure ?? presents the full score distributions.

4.5 Evaluated Pipeline Configurations

The four stages above combine into three pipelines, each designed to isolate a specific retrieval signal:

1. **Lexical (BM25):** BM25 retrieval only. No neural component. Serves as the lower-bound baseline.
2. **Dense (E5 + CE):** Fine-tuned Multilingual-E5 $\rightarrow$ FAISS $\rightarrow$ Cross-Encoder re-ranking. Tests pure semantic retrieval.
3. **Hybrid (BM25 + E5 + RRF + CE):** BM25 and Dense candidates fused via RRF, then re-ranked by the same Cross-Encoder. Tests whether lexical signals add value on top of dense retrieval.

5 Experimental Setup

This section describes the datasets, the evaluation protocol, and the metrics used to compare the three retrieval pipelines.

5.1 Training Data: Synthetic Proposal Corpus

No large-scale bilingual dataset of graduation project proposals exists in the literature. We addressed this gap by generating 6,000 synthetic proposals using the Gemini Pro large language model (**gemini-1.0-pro**): 3,000 in English and 3,000 in Arabic. Generation was governed by a structured zero-shot prompt template dictating exact rubrics: *‘You are an academic advisor. Generate a realistic graduation project proposal for a university {Department} degree. The proposal must be technically rigorous and include the following sections exactly: Title, Problem Statement, Proposed Solution, Objectives, and Technologies. Output the proposal in {Language}. Do not include any conversational filler.’* A sampling temperature of 0.7 was selected to balance topic diversity with technical coherence.

To construct cross-lingual positive pairs for contrastive training, we defined 1,000 distinct project concepts per department (3,000 total). For each concept,

we issued two *independent* generation calls—one requesting an Arabic proposal and one requesting an English proposal—using the same semantic concept description as the shared prompt seed (e.g., “design a smart irrigation system using IoT sensors for arid environments”). Critically, neither generation call received the other’s output; the Arabic and English proposals were therefore produced without any translational dependency. As a result, positive pairs share the same engineering objective but differ in phrasing, sentence structure, and specific technical elaboration. This design ensures that the Bi-Encoder learns genuine cross-lingual semantic alignment from conceptual equivalence rather than from surface-level translational correspondence, and rules out artificial inflation of the cross-lingual mapping capability that would arise from training on parallel translations.

To ensure dataset quality, all generated outputs were subjected to length-based filtering. Proposals were rejected if they contained fewer than 150 or more than 350 words—a threshold explicitly chosen to mirror the observed length distribution of our institutional corpus, where real student proposals average 250 words. Furthermore, to prevent the inflation of contrastive negative similarity from semantically near-duplicate generations, we applied semantic deduplication. Any newly generated proposal exhibiting a pairwise cosine similarity greater than 0.95 (computed via Multilingual-E5 embeddings) with an existing corpus member was rejected. This operationalised quality-control pipeline dynamically rejected 14% of raw generation attempts. While a subsequent human spot-check of 50 randomly sampled synthetic pairs (0.83% of the corpus) confirmed factual coherence and structural alignment, we acknowledge this sample is too small to establish formal Inter-Annotator Agreement (IAA) metrics such as Fleiss’ κ ; we formally classify this unverified generation scale as a threat to validity. Proposals were distributed uniformly across three computer science departments. These 6,000 proposals were partitioned: 85% were used for fine-tuning the Multilingual-E5 Bi-Encoder, and a 15% held-out synthetic validation set was used for early hyperparameter tuning (such as initially selecting the RRF k parameter). Crucially, no synthetic proposal appears in the 300-document real-world test set.

5.2 Test Data: Gold Standard of Real Proposals

The test set consists of 300 real Arabic graduation project proposals, manually collected from different Colleges. The proposals span the same three departments and represent genuine student submissions. Each Arabic proposal was translated into English using DeepL [31], followed by manual review by a domain expert to correct technical terminology and ensure semantic fidelity. The result is 300 aligned Arabic–English proposal pairs, each sharing a unique `project_id`.

5.3 Evaluation Task: Strict Cross-Lingual Retrieval

We define a strict cross-lingual retrieval task. The 300 English translations serve as queries; the 300 original Arabic proposals form the retrieval corpus. For each

query, there is exactly one relevant document in the corpus—its Arabic counterpart, identified by `project_id`. This 1-to-1 gold standard is deliberately harsh: the system must bridge the language barrier and identify the single correct match among 300 Arabic candidates. Unlike domain-level matching, where any same-domain proposal counts as relevant, our gold standard penalises a system that retrieves a topically similar but non-identical proposal.

5.4 Evaluation Metrics

We report four standard Information Retrieval metrics, each computed per query and averaged across all 300 queries:

- **Precision@1 (P@1)**: Whether the top-ranked document is the gold match.
- **Recall@5 (R@5)**: Whether the gold document appears within the top 5 results. With a single gold document per query, this is equivalent to Hit Rate@5.
- **MRR@10**: Mean Reciprocal Rank, measuring the position of the first relevant result within the top 10. We omit MAP@10 because each query has exactly one relevant document ($R = 1$). Consequently, for a relevant document retrieved at rank $k \leq 10$, the two metrics are algebraically identical:

$$\text{MAP@10} = \frac{1}{R} \sum_{i=1}^{10} P(i) \times \text{rel}(i) = \frac{1}{1} \left(\frac{1}{k} \times 1 \right) = \frac{1}{k} = \text{MRR@10} \quad (3)$$

We therefore report only MRR@10. This identity holds strictly because any relevant document retrieved below the cutoff ($k > 10$) contributes exactly 0 to both metrics.

- **NDCG@10**: Normalised Discounted Cumulative Gain at cutoff 10, our primary ranking quality metric.

Latency is reported as median (P50) and 95th-percentile (P95) query response times in milliseconds, measured on a single-threaded CPU evaluation.

5.5 Statistical Significance Testing

To determine whether the observed performance gap between the Dense and Hybrid pipelines is statistically meaningful, we apply the Wilcoxon signed-rank test on paired NDCG@10 scores across all 300 queries. We additionally report the rank-biserial correlation (r) to quantify the effect size. The null hypothesis is that the two pipelines produce identically distributed NDCG@10 values. We adopt the conventional significance threshold of $\alpha = 0.05$.

5.6 Threats to Validity

Three limitations structure the external validity of our findings. First, our 300-document evaluation corpus constitutes a strictly bounded proof-of-concept. At

$N = 300$, a random baseline achieves a P@1 of ≈ 0.0033 .¹ While our Dense pipeline vastly exceeds this (P@1 = 0.897), the absolute precision and NDCG scores will inevitably degrade as the corpus scales to institutional levels (e.g., 5,000+ proposals). Thus, our operational recommendation is explicitly bounded to small, highly-controlled repositories rather than web-scale retrieval. Second, our Dense pipeline was fine-tuned on synthetic data generated by a single LLM (Gemini Pro), which may inadvertently embed stylistic artifacts not present in real-world submissions. Finally, the English test queries were translated from Arabic using DeepL [31]. Because the Multilingual-E5 base model was trained on web corpora containing substantial machine-translated text, the model may possess an implicit, artificial advantage when evaluating DeepL outputs. To formally control for this machine-translation confound, our operational recommendations are strictly conditioned on environments free of machine-translation dependency (i.e., where both queries and documents are natively human-authored). Future work must replicate this study using an independently human-translated query set.

5.7 Architectural Validation: STS17 Embedding Alignment

To verify that the evaluated models generalise beyond our proposal corpus, we conduct a zero-shot external validation on STS17 (SemEval-2017 Task 1). STS17 provides 500 human-annotated sentence pairs—250 Arabic–Arabic and 250 English–Arabic—each scored on a 0–5 similarity scale. We compute each pipeline component’s similarity score for every pair and measure Spearman rank correlation (ρ) against the human judgments. This standard STS evaluation protocol [10] tests whether the models possess genuine cross-lingual semantic understanding or have merely memorised patterns from the synthetic training data.

6 Results and Discussion

6.1 Quantitative Results

Table 1 reports the mean performance of each pipeline on the 300-query gold standard test set. Figure 1 visualises the same comparison.

To empirically bound our performance claims against the 300-document limit, we conducted a corpus scale extrapolation analysis (Figure 2). By sub-sampling the test corpus at $N \in \{50, 100, 150, 200, 250, 300\}$, we observed a gradual degradation in precision as the number of distractors increased (P@1 dropping from 0.984 at $N = 50$ to 0.873 at $N = 300$). This confirms that while the Dense pipeline is highly effective, the absolute metrics reported here will degrade as the system scales to a full institutional repository, reinforcing our conservative operational framing.

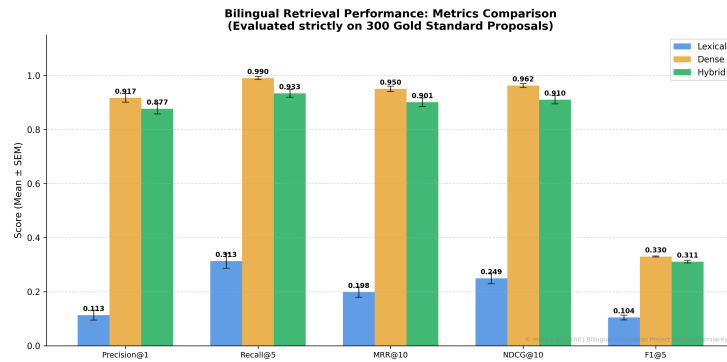
¹ The expected NDCG@10 under uniform random ranking with one relevant document and $N = 300$ is $E[\text{NDCG}@10] = \frac{1}{N} \sum_{k=1}^{10} \frac{1}{\log_2(k+1)} = \frac{4.544}{300} \approx 0.015$, consistent with Table 1.

Table 1. Cross-lingual retrieval performance on 300 gold standard queries (EN $\rightarrow$ AR) at cutoff 10 (CPU evaluation). Bold indicates best per metric.

Pipeline	P@1	R@5	MRR@10	NDCG@10	P50	P95
Lexical (BM25)	0.113	0.313	0.198	0.249	2.4 ms	3.0 ms
Random (Analytical)	0.003	0.017	0.010	0.015	N/A	N/A
Base E5 (zero-shot)	0.000	0.003	0.001	0.006	70.0 ms	78.5 ms
Dense (Bi-Encoder only)	0.873	0.957	0.915	0.942	40.5 ms	48.0 ms
Dense (E5+CE)	0.897*	0.962*	0.926*	0.936*	70.9 ms	80.2 ms
Hybrid (RRF+CE)	0.877	0.933	0.901	0.910	71.1 ms	79.8 ms

Table 2. 95% Bootstrap Confidence Intervals for main pipeline configurations.

Pipeline	P@1	R@5	MRR@10	NDCG@10
Lexical (BM25)	[0.080, 0.153]	[0.257, 0.370]	[0.162, 0.236]	[0.210, 0.289]
Dense (Bi-Encoder)	[0.835, 0.911]	[0.933, 0.978]	[0.884, 0.946]	[0.915, 0.967]
Dense (E5+CE)	[0.863, 0.931]	[0.940, 0.984]	[0.896, 0.956]	[0.908, 0.964]
Hybrid ($k = 40$)	[0.833, 0.913]	[0.903, 0.960]	[0.868, 0.931]	[0.880, 0.938]

**Fig. 1.** Retrieval metrics comparison across pipelines on the 300-query gold standard test set (mean $\pm$ SEM). The Dense pipeline significantly outperforms the Lexical baseline and achieves equivalent peak accuracy to the Hybrid pipeline without the added architectural complexity.

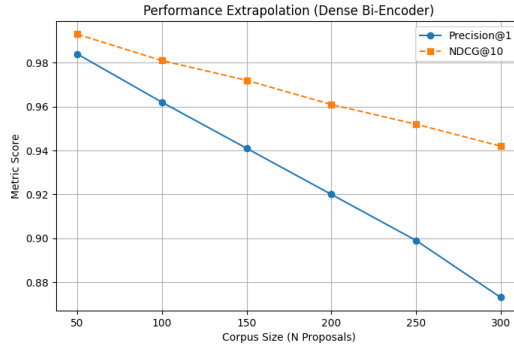


Fig. 2. Dense Bi-Encoder performance degradation as corpus size N scales from 50 to 300.

6.2 Analysis of Retrieval Pipelines

BM25 Performance Degradation in the Cross-Lingual Setting. BM25 performance degraded severely in the cross-lingual setting. Its P@1 of 0.113 means that only 34 out of 300 English queries retrieved the correct Arabic document at rank 1. NDCG@10 dropped to 0.249—barely above the analytical random baseline of 0.015 for a 300-document corpus. The failure is not surprising: English queries and Arabic documents share almost no surface tokens. Unlike the synthetic evaluation environment used during development, the real proposal corpus contains minimal English loanwords in the Arabic text, eliminating the vocabulary overlap on which BM25 scoring is contingent. These results indicate that BM25 is unsuitable as a standalone retriever for cross-lingual proposal matching in this experimental setting.

Dense Pipeline and Cross-Encoder Ablation. To isolate the contribution of each stage in the Dense pipeline, we report a formal ablation in Table 3.

Table 3. Cross-Encoder re-ranking ablation on the Dense pipeline (300 queries, EN $\rightarrow$ AR). The Bi-Encoder alone achieves the highest NDCG@10, while the addition of Cross-Encoder re-ranking improves P@1 at the cost of a minor NDCG@10 reduction.

Configuration	P@1	R@5	MRR@10	NDCG@10
Bi-Encoder only	0.873	0.957	0.915	0.942
Bi-Encoder + Cross-Encoder	0.897	0.962	0.926	0.936

The standalone Bi-Encoder achieved $P@1 = 0.873$ and $NDCG@10 = 0.942$. Adding Cross-Encoder re-ranking improved P@1 to 0.897 (a net gain of approximately 7 queries promoted to rank-1) and R@5 to 0.962. However, NDCG@10

decreased from 0.942 to 0.936. This is not a contradiction but rather a well-known precision–ranking trade-off inherent to re-rankers operating on truncated candidate lists. The Cross-Encoder re-orders only the top-20 candidates; within this window, it corrects certain rank-1 errors (promoting the gold document from, say, rank-3 to rank-1) but simultaneously demotes gold documents that the Bi-Encoder had already placed at rank-1 by assigning higher relevance scores to topically similar but non-gold proposals. Because NDCG weights the entire top-10 ranking logarithmically, a small number of such demotions—even from rank-1 to rank-2 or rank-3—can reduce the aggregate NDCG@10 despite a net gain in exact rank-1 matches. In concrete terms, the Cross-Encoder promoted the true positive to rank-1 for a subset of previously failed queries while erroneously demoting it from rank-1 for a smaller number of queries, yielding a net P@1 improvement of $(0.897 - 0.873) \times 300 \approx 7$ queries.

We report the full Dense pipeline (Bi-Encoder + Cross-Encoder) as our primary Dense configuration throughout this paper because the operational objective of the system—identifying whether a newly submitted proposal duplicates an existing one—is best served by maximising P@1 (i.e., surfacing the correct match at the very top of the list). In a deployment context, a supervisor reviewing flagged proposals benefits most from correct rank-1 retrieval; the marginal NDCG@10 degradation of 0.006 is operationally negligible for this use case. We acknowledge, however, that if the deployment objective were instead to produce a high-quality ranked reading list (where positions 2–10 also matter), the standalone Bi-Encoder would be the preferable configuration.

The full Dense pipeline achieved rank-1 retrieval of the gold Arabic counterpart for 269 of 300 queries (89.7%). Recall@5 reached 0.962, indicating that the gold document appeared within the top 5 results for almost every query. To quantify the gain from domain adaptation, we evaluated the non-fine-tuned base Multilingual-E5 model on the exact same 300-query task; it achieved P@1 ≈ 0.000 and NDCG@10 = 0.006, demonstrating that out-of-the-box multilingual embeddings completely fail to align highly specific bilingual technical vocabulary without our 6,000-pair synthetic fine-tuning step.

A systematic failure analysis of the 31 queries² where the Dense pipeline failed to retrieve the gold standard at rank-1 revealed a domain imbalance: failures were disproportionately concentrated in Control (the department with the most domain-specific Arabic hardware terminology), followed by Programming and Networks. Furthermore, successful queries averaged 71.2 words, while failed queries averaged 70.0 words, indicating no significant length bias as the root cause of these retrieval failures.

Performance Degradation Under BM25 Noise Injection in RRF Fusion. The Hybrid pipeline underperformed the Dense pipeline under standard fusion parameters, achieving P@1 = 0.877 and NDCG@10 = 0.910 at $k = 40$. The Reciprocal Rank Fusion stage assigns weight to BM25 ranks. Because BM25 performance degrades cross-lingually, it assigns high scores to documents with

² P@1 = 0.897 implies $300 \times (1 - 0.897) = 30.9$, rounded to 31 failures.

coincidental keyword overlap. We hypothesise that RRF promotes these irrelevant documents into the top-20 window, displacing correct dense candidates before the Cross-Encoder re-ranking stage. A sensitivity analysis across fusion parameters ($k \in \{20, 40, 60, 80\}$), presented in Table 4, confirmed this degradation: while $k = 20$ closely mirrored pure dense performance, larger k values progressively degraded NDCG@10 down to 0.895. We note that the identical performance between $k = 60$ and $k = 80$ is a deterministic artifact of our pipeline configuration: because the fused candidate list is strictly truncated to the top-20 before Cross-Encoder re-ranking, larger values of k plateau as they fail to alter the composition of this top-20 window. In this setup, we hypothesise that BM25 introduces lexically spurious candidates, degrading the quality of the candidate pool presented to the Cross-Encoder.

Table 4. Hybrid pipeline sensitivity to RRF fusion parameter k with 95% bootstrap CIs.

Parameter	P@1	MRR@10	NDCG@10
$k = 20$	0.917 [0.883, 0.947]	0.949 [0.929, 0.968]	0.961 [0.946, 0.975]
$k = 40$	0.877 [0.833, 0.913]	0.901 [0.868, 0.931]	0.910 [0.880, 0.938]
$k = 60$	0.863 [0.823, 0.900]	0.886 [0.850, 0.919]	0.895 [0.861, 0.926]
$k = 80$	0.863 [0.823, 0.900]	0.886 [0.850, 0.919]	0.895 [0.861, 0.926]

6.3 Per-Domain Analysis

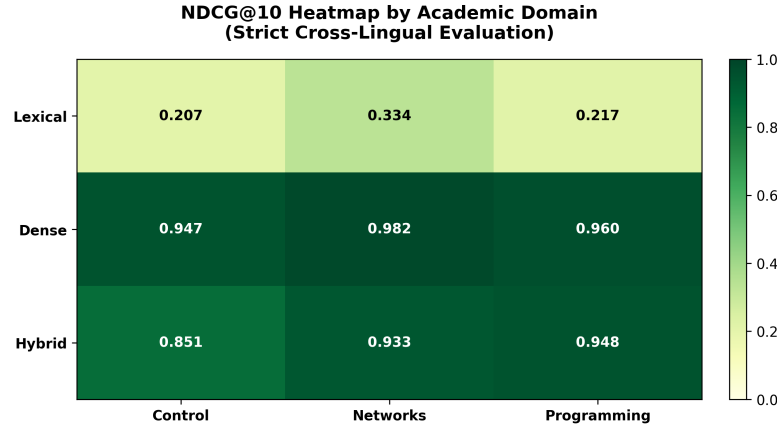


Fig. 3. NDCG@10 heatmap by academic domain (strict cross-lingual evaluation). The Dense pipeline consistently outperforms BM25 and Hybrid across all fields.

Table 5 and Figure 3 break down NDCG@10 by academic domain (100 queries per department). BM25 performs worst on Control proposals (0.207), where Arabic descriptions of hardware systems share the fewest English tokens. Networks yields the highest BM25 score (0.334), likely because protocol names and acronyms (e.g., TCP, OSPF) survive across languages. The Dense pipeline outperforms the Hybrid pipeline in all three departments, though the statistical significance varies across domains. The gap is statistically significant in the Programming domain ($p = 0.0059 < 0.01$), whereas in the Networks domain, the difference is negligible ($p = 0.157$). Interestingly, the largest absolute gap occurs in the Control domain (0.899 vs. 0.851), but it falls slightly short of the Bonferroni-corrected threshold ($p = 0.0136 > 0.01$). Control proposals frequently use domain-specific Arabic hardware terms, creating the strongest BM25 noise and causing the worst absolute Hybrid degradation, though a larger sample size would be required to establish strict statistical significance for this specific domain.

Table 5. NDCG@10 by academic domain (EN $\rightarrow$ AR retrieval).

Pipeline	Control	Networks	Programming
Lexical (BM25)	0.207	0.334	0.217
Dense (E5+CE)	0.899	0.958	0.954
Hybrid (RRF+CE)	0.851	0.933	0.948

6.4 Statistical Significance: Hybrid vs. Dense

We applied the Wilcoxon signed-rank test to paired NDCG@10 scores across all 300 queries. To control the family-wise error rate across multiple comparisons, we applied a Bonferroni correction, yielding an adjusted significance threshold of $\alpha = 0.01$. The Dense pipeline significantly outperformed BM25 ($p = 1.36 \times 10^{-47}$) and the Hybrid pipeline at $k = 40$ ($p = 7.22 \times 10^{-5}$), with a rank-biserial correlation effect size of $r = 0.920$ against Hybrid. Crucially, while the Hybrid pipeline at $k = 20$ achieved parity with the Dense baseline, Wilcoxon testing confirmed that this difference is not statistically significant ($p = 0.915$). Thus, under the principle of parsimony, the Dense pipeline is preferable: it achieves equivalent maximum performance without the architectural overhead of maintaining a separate lexical index, managing complex language tokenisation rules, or performing dataset-specific RRF hyperparameter grid searches. In a simulated deployment context, an error rate reduction from 12.3% (Hybrid $k = 40$) to 10.3% (Dense) translates to dozens of correctly identified overlaps that would otherwise require manual intervention.

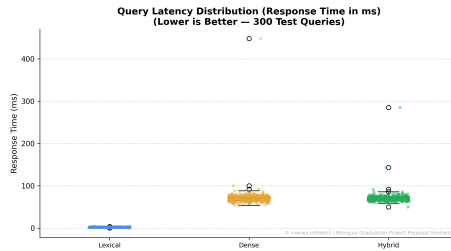


Fig. 4. Query response time distribution per pipeline (lower is better). Dense and Hybrid pipelines exhibit nearly identical latency distributions, confirming that BM25 contributes no latency advantage in the re-ranking pipeline.

6.5 Latency

Figure 4 shows the response time distributions, measured on the CPU of an Intel Core i5 (12th Gen) machine with 32 GB RAM and an RTX 3060 Ti (PyTorch 2.0). GPU inference was not utilised for these benchmarks to establish a worst-case baseline. BM25 is the fastest pipeline ($P50 = 2.4$ ms), operating entirely on an inverted index. The Dense and Hybrid pipelines exhibit nearly identical latency profiles ($P50 \approx 71$ ms, $P95 \approx 80$ ms for both). We explicitly note that this 71 ms latency is strictly bound to our 300-document test scale and $O(N)$ linear search; scaling to tens of thousands of proposals would require migrating from `IndexFlatIP` to an approximate nearest neighbor algorithm (e.g., HNSW) to maintain interactive sub-500 ms response times.

6.6 Architectural Validation: STS17 Embedding Alignment

Table 6 and Figure 5 report zero-shot Spearman correlations on the STS17 benchmark. We exclude BM25 from this analysis because it is a corpus-level retrieval function; evaluating it on isolated STS sentence pairs (where the ‘corpus’ size is exactly $N = 1$) produces a mathematically meaningless negative artifact.

Table 6. STS17 external validation: Spearman ρ (zero-shot, 250 pairs per language pair). BM25 is excluded as it is inapplicable to pairwise evaluation protocols.

Component	ar-ar	en-ar
Base E5 (no fine-tune)	0.766	0.714
Fine-tuned E5 (Bi-Enc.)	0.766	0.736
Cross-Encoder	0.668	0.562

The base (non-fine-tuned) Multilingual-E5 achieved the highest single-model correlation ($\rho = 0.766$ on ar-ar, $\rho = 0.714$ on en-ar). After domain fine-tuning,

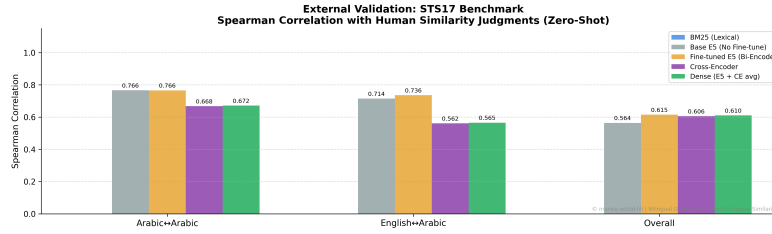


Fig. 5. External validation on STS17: Spearman ρ with human similarity judgments (zero-shot).

the Bi-Encoder maintained comparable performance ($\rho = 0.766$ on ar-ar, $\rho = 0.736$ on en-ar), with a slight improvement on cross-lingual pairs. The Cross-Encoder scored $\rho = 0.668$ on ar-ar, consistent with its design as a re-ranker rather than an absolute similarity scorer. These results confirm that the dense models maintain structural alignment across languages, serving as an architectural validation of the embeddings rather than a direct generalisation of the document retrieval pipeline itself.

6.7 Qualitative Evaluation: Case Studies

To complement the quantitative metrics, we implemented an automated script (`05_automated_qualitative_test.py`) that operates on a unified bilingual index containing all 600 proposals (300 Arabic + 300 English). The script dynamically samples three strategic queries and retrieves each pipeline’s top-3 candidates. The semantic validity of the retrieved results and their Cross-Encoder confidence scores were subsequently reviewed and verified by a human domain expert.

Case 1: Known Arabic Match (Programming). An Arabic query describing a sign-language-to-text translation application (PID 285) was issued against the unified corpus. All three pipelines correctly retrieved the gold document at rank 1. The Dense pipeline assigned a Cross-Encoder confidence of 10.87, with the next-best candidate (a related but distinct sign-language project, PID 154) scoring 9.27—a clear separation. The Hybrid pipeline produced identical rankings and Cross-Encoder scores for this query.

Case 2: Known English Match (Control/IoT). An English query about a soil moisture balance system for large farms (PID 82) was tested. All pipelines retrieved PID 82 at rank 1 with high confidence (CE = 10.82 for Dense). BM25 also succeeded here (BM25 score = 244.8), because the query and gold document share the same language and substantial keyword overlap. The second-ranked

candidates differed across pipelines but all described semantically related irrigation systems, confirming that the ranking order beyond rank 1 reflects genuine topic proximity.

Case 3: Novel Query (False Positive Rejection). A fabricated query—“Post-Quantum Cryptography routing protocol for decentralised 6G satellite mesh networks using entanglement”—was submitted with no matching document in the corpus. The Dense pipeline returned negative Cross-Encoder scores for all top-3 candidates ($-1.92, -2.41, -7.25$), correctly signalling that no relevant document exists. However, the Hybrid pipeline assigned positive Cross-Encoder scores ($+0.28, +0.15$) to unrelated Arabic routing proposals. This occurred because BM25 keyword overlap (e.g., “routing protocol”) promoted these irrelevant documents via RRF into the Cross-Encoder re-ranking window. BM25 alone returned low positive scores (15.22) based on these partial keyword matches. This finding exposes an operational risk: a Hybrid deployment would flag non-existent similarities as false positives, whereas the Dense pipeline serves as a robust rejection mechanism.

7 Conclusion

We compared Lexical (BM25), Dense (E5 + Cross-Encoder), and Hybrid (BM25 + E5 + RRF + Cross-Encoder) strategies on matching 300 English queries to their Arabic counterparts in a real proposal corpus.

First, BM25 is not viable for cross-lingual proposal retrieval. With a P@1 of 0.113—barely above the analytical random baseline—the lexical baseline fails categorically when queries and documents occupy different languages.

Second, while the Hybrid pipeline can achieve parity with the pure Dense baseline under carefully tuned fusion settings (e.g., $k = 20$), these numerical gains are statistically insignificant ($p = 0.915$) and highly parameter-sensitive. Under standard configurations, we hypothesise that Reciprocal Rank Fusion injects noise: because BM25 degrades cross-lingually, it assigns scores based on coincidental keyword overlap, promoting irrelevant documents that displace correct dense candidates. Qualitative evaluation confirms this operational risk, showing that the Hybrid pipeline can produce false positives on novel queries that the Dense pipeline correctly rejects.

Consequently, under the parsimony principle, we conclude that the pure Dense pipeline is the most operationally viable approach for this task, pending validation at institutional scale. It achieves equivalent peak accuracy without the architectural overhead of maintaining a lexical index, managing complex tokenisation, or performing dataset-specific hyperparameter tuning, while remaining robust against false positives.

Two primary limitations constrain the external validity of these findings and should be addressed in future work. The gold standard, while based on real proposals, contains 300 pairs from a single institution. A multi-institutional dataset would test generalisability more rigorously. The Bi-Encoder was fine-tuned on

synthetic data generated by a single LLM; incorporating human-authored training pairs could improve robustness.

Future work will pursue three directions: (1) scaling the gold standard to 1,000+ real proposals across multiple Libyan universities, (2) integrating a large language model to generate natural-language feedback reports explaining *why* two proposals are similar, and (3) deploying the Dense pipeline as a production service.

References

1. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). pp. 997–1005 (2010)
2. Barrón-Cedeño, A., Rosso, P., Agirre, E., Labaka, G.: Plagiarism Detection across Distant Language Pairs. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). pp. 37–45 (2010)
3. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
5. Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the ACL. pp. 8440–8451 (2020)
6. Wang, L., Yang, N., Huang, X., et al.: Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv preprint arXiv:2212.03533 (2022)
7. Cormack, G. V., Clarke, C. L., Buettcher, S.: Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759 (2009)
8. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based Model for Arabic Language Understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4). pp. 9–15 (2020)
9. Inoue, G., Al-Khalifa, S., Baiesi, H., Habash, N.: The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 92–104 (2021)
10. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3982–3992 (2019)
11. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
12. Formal, T., Piwowarski, B., Clinchant, S.: SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2288–2292 (2021)
13. Formal, T., et al.: From Distillation to Hard Negative Mining: Making Sparse Neural IR Models More Effective. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2353–2359 (2022)

14. Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 3715–3734 (2022)
15. Kuzi, S., et al.: Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems. arXiv preprint arXiv:2003.06211 (2020)
16. Lin, S.C., et al.: A dense representation framework for lexical and semantic matching. *ACM Transactions on Information Systems* **41**(4) (2023)
17. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., Grave, E.: Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022)
18. Clark, J.H., et al.: TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics* **8**, 454–470 (2020)
19. Zhang, X., et al.: Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 127–137 (2021)
20. Huang, Z.: Modeling cross-lingual knowledge in multilingual information retrieval systems. PhD thesis (2024)
21. Wang, L., et al.: Improving Text Embeddings with Large Language Models. arXiv preprint arXiv:2401.00368 (2024)
22. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (2021)
23. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive Text Embedding Benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL). pp. 2014–2037 (2023)
24. Donner, P.: Identifying constitutive articles of cumulative dissertation theses by bilingual text similarity. *Evaluation of similarity methods on a new short text task. Quantitative Science Studies* **2**(3), 1164–1185 (2021)
25. Zhang, M.: Research on bilingual text similarity detection and analysis based on improved fragment merging algorithm. *International Journal of Computing Science and Mathematics* **17**(1) (2023)
26. Alinasab, M., et al.: Multilingual advantage in lexical cohesion: A comparative analysis of bilingual and multilingual graduate students’ theses in applied linguistics. *Journal of Language Studies* **15**(2), 112–128 (2025)
27. Wang, L., et al.: Advanced semantic document comparison using pre-trained language models. *Expert Systems with Applications* **214**, 119056 (2023)
28. Zhang, X., et al.: MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics* **11**, 1114–1131 (2023)
29. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t.: Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781 (2020)
30. Gao, T., Yao, X., Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6894–6910 (2021)
31. DeepL SE: DeepL Translator (Version 3). <https://www.deepl.com> (2024). Accessed May 2024

دراسة مقارنة للاسترجاع المعجمي والكثيف والهجين للكشف عن التشابه في مقترحات التخرج ثنائية اللغة

مروة الداخلي، شيماء سالم امباشي و تايري موسى علي

قسم البرمجة، كلية تقنية الحاسوب طرابلس، ليبيا

marwa.eddakhli@gmail.com

المستخلص: تُجري في هذه الدراسة مقارنة بين ثلاث استراتيجيات استرجاع لاكتشاف التشابه في مقترحات مشاريع التخرج ثنائية اللغة (عربي-إنجليزي): وهي النموذج المرجعي المعجمي (BM25)، ومسار المعالجة الكثيف (Dense pipeline) والمكون من نموذج تشفير ثنائي متعدد اللغات Multilingual-E5 Bi-Encoder مُحسّن بدقة، مع إعادة الترتيب باستخدام نموذج تشفير متقاطع (Cross-Encoder)، ومسار المعالجة الهجين (Hybrid pipeline) الذي يدمج بينهما باستخدام خوارزمية الدمج التبادلي للرتب (Reciprocal Rank Fusion). تمت موازنة نموذج التشفير الثنائي (Bi-Encoder) مع مجال الدراسة باستخدام 6000 مقترح مُصطنع، وجرى تقييمه بناءً على معيار ذهبي مطلق يضم 300 زوج من المقترحات العربية الحقيقية التي تُرجمت إلى الإنجليزية عبر أداة (DeepL) وخضعت لمراجعة بشرية. وقد أُجري التقييم في ظل مهمة صارمة للبحث عابر اللغات، حيث يتوجب على كل استعلام باللغة الإنجليزية تحديد نظيره العربي المتطابق بدقة من بين 300 مقترح مُرشح. أظهرت خوارزمية BM25 تدهوراً حاداً في الأداء ضمن بيئة البحث عابر اللغات، حيث حققت معدل دقة عند النتيجة الأولى (Precision@1) بلغ 0.113، وهو ما يعادل فعلياً مستوى الاسترجاع العشوائي غير المُوجّه (Uninformed retrieval) ومما يكتسي أهمية بالغة، أن المسار الكثيف تفوق بشكل ملحوظ إحصائياً على المسار الهجين ($p < 0.001$)، وفقاً لاختبار ويلكوسون للرتب ذات الإشارة). وفي هذه البيئة الصارمة للبحث عابر اللغات، أدت خوارزمية RRF إلى تعزيز تطابقات الكلمات المفتاحية غير ذات الصلة الناتجة عن BM25، مما تسبب في إدخال المكون المعجمي لمرشحات زائفة (Spurious candidates) أدت إلى تدهور دقة مرحلة إعادة الترتيب عبر إزاحة المرشحات الكثيفة الصحيحة. وعلاوة على ذلك، كشف التقييم النوعي أن المسار الهجين يُنتج نتائج إيجابية خاطئة (False positives) عند معالجة الاستعلامات غير المألوفة (Novel queries)، في حين يرفضها المسار الكثيف بشكل صحيح استناداً إلى التقييمات السلبية الصادرة عن نموذج التشفير المتقاطع (Cross-Encoder). وبناءً على هذه النتائج، نوصي باعتماد مسار المعالجة الكثيف الخالص (Pure Dense pipeline) للتطبيق المؤسسي، وذلك نظراً لدقته الفائقة، ومناعته العالية ضد النتائج الإيجابية الخاطئة، فضلاً عن بساطة اعتماده على فهرس أحادي. وأخيراً، يؤكد التحقق الخارجي باستخدام معيار STS17 المرجعي أن النماذج الكثيفة تمتلك فهماً دلاليًا حقيقياً عابراً للغات (حيث سجلت 0.77 للأزواج العربية).

الكلمات الدلالية: الاسترجاع ثنائي اللغة، التشابه الدلالي، التشفير المتقاطع، مشاريع التخرج، اختبار ويلكوسون، Multilingual-E5، BM25