



DEAM: Diverse Embeddings for Arabic Multimodal Retrieval

Omer Nacar ^{1*}, Fatimah Mohamed Emad Elden ², Rahaf Saeed Al Hallay ³, Mona A. Azim ⁴, Aya Hamod ⁵, Mohamed Zaytoon ³, Nour Qandos ³, Israa Elhosiny ⁶, Mohamed Abdelaziz ⁵, Yousra Alaa El-Ghawi ³, Serry Sibae ⁷, Hadi Hamoud ⁸, Norah Naji Al-Zaid ⁹, Muhammed Ragab ¹⁰, Khlood Al Jallad ¹¹

¹Tuwaiq Academy, ²Trouve Labs, ³NAMAA Community, ⁴Ain Shams University, ⁵Global Brands Group, ⁶Alexandria University, ⁷Prince Sultan University, ⁸Arab Center for Research and Policy Studies, ⁹King Khalid University, ¹⁰Leibniz-Institute for Educational Media, Georg-Eckert-Institute, ¹¹Arab International University

o.najar@tuwaiq.edu.sa



Figure 1: Overview of DEAM: nine Arabic document domains spanning visual-rich pages (charts and tables) and prose-rich sources (Wikipedia, books, and government documents), with held-out queries validated by 21 Arabic annotators.

Abstract

Modern visual document retrieval (VDR) models remain strongly English-centric, leaving Arabic substantially under-served in this rapidly advancing field. To address this gap, we introduce DEAM, the first open Arabic visual document retrieval dataset explicitly designed for fine-tuning and evaluating multimodal retrievers on Arabic. DEAM is built through a human-in-the-loop construction pipeline with **21 community** annotators, combining Arabic document rendering, query creation, validation, and retrieval-oriented hard-negative min-

ing. The dataset comprises **18,275** rendered Arabic documents spanning nine source domains and **82,727** contrastive retrieval triplets, each including five mined hard negatives and one easy negative for contrastive fine-tuning. Zero-shot evaluation of 27 retrievers shows that multilingual text retrievers outperform visual retrievers on Arabic by 13.6 nDCG@10, except on layout-rich news pages. DEAM establishes a public foundation for Arabic visual document retrieval.¹

¹<https://huggingface.co/collections/NAMAA-Space/deam-for-arabic-multimodal-retrieval>

1 Introduction

Recent multimodal embedding models have transformed information retrieval. Encoders such as Qwen3-VL-Embedding (Li et al., 2026) and Gemini Embedding (Lee et al., 2025) map text, images, document pages, and video into shared semantic spaces for retrieval, semantic textual similarity, clustering, and classification. On the visual document side, late-interaction encoders such as ColPali (Faysse et al., 2025), the ColQwen family (Faysse et al., 2024), DSE (Ma et al., 2024), and unified frameworks such as VLM2Vec-V2 (Meng et al., 2025) directly index rendered document pages and achieve strong results on standard benchmarks. However, visual retrievers operate on page images and must recover text from pixels, whereas conventional text retrievers access clean textual content; their comparison is therefore practically useful but not a fully controlled modality comparison.

These advances have been measured primarily on English data, and their gains do not transfer uniformly across languages. Recent multilingual visual retrieval benchmarks show that state-of-the-art visual embedders lose substantial accuracy on non-English scripts (Osmulski et al., 2025). Despite Arabic being used by more than 400 million people, no open Arabic visual document retrieval (VDR) resource pairs rendered documents with human-validated queries. The community, therefore, lacks a dedicated corpus for training and evaluating Arabic visual retrieval models.

Arabic introduces challenges that are easy to underestimate. Its script is bidirectional, contextually shaped, and may include optional diacritics. Rendering pipelines that do not fully support Arabic can produce detached or mis-shaped glyphs, corrupting document images and their retrieval signals. Arabic also includes substantial variation between Modern Standard Arabic (MSA) and regional dialects. DEAM focuses on MSA; dialectal and regional varieties remain an important direction for future benchmark extensions.

We address this gap with DEAM, the first open Arabic visual document retrieval resource. Unlike recent Arabic OCR and instruction-tuning resources, DEAM targets retrieval over rendered documents. Our contributions are:

- **A multi-domain Arabic VDR corpus** spanning nine source categories: four visual-heavy categories (tables; bar, line, and pie charts) and five prose-heavy categories (Wikipedia, news, books,

science, and government), rendered through an HTML+Playwright pipeline that preserves Arabic shaping and bidirectional layout.

- **A reproducible three-source query-generation pipeline** combining vision-capable and text-only LLMs to produce diverse Arabic queries, followed by vision-based refinement into explicitly typed *direct*, *comparative*, and *long-tail* intents.
- **A community human-in-the-loop layer** that produces a human-validated dataset under a two-of-three consensus protocol on Argilla (Argilla, 2024), including a separate verdict on rendered-image quality.
- **Ready-to-use contrastive triplets and reproducible tooling** with mined hard negatives from ColQwen2.5 (Faysse et al., 2024) and TF-IDF, together with evaluation and fine-tuning harnesses.

2 Related Work

Multimodal and Text Embedding Models.

Modern retrieval is dominated by large multilingual embedding encoders such as BGE-M3 (Chen et al., 2024), multilingual-E5 (Wang et al., 2022, 2024), GTE-Qwen2 (Li et al., 2023), Qwen3-Embedding (Zhang et al., 2025), and Gemini Embedding (Lee et al., 2025) on the text side, with CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and the Qwen3-VL-Embedding (Li et al., 2026) and VLM2Vec-V2 (Meng et al., 2025) families extending the paradigm into unified visual-text encoders. All claim multilingual coverage but none is trained or benchmarked on Arabic VDR.

Visual Document Retrieval.

Document retrieval has shifted from OCR-then-text pipelines (BM25 (Robertson and Zaragoza, 2009), DPR (Karpukhin et al., 2020), ColBERT (Khattab and Zaharia, 2020)) to vision-aware retrievers that encode rendered pages directly: DSE (Ma et al., 2024) as a single dense vector, and ColPali (Faysse et al., 2025) and ColQwen2.5 (Faysse et al., 2024) via late-interaction MaxSim. These benchmarks are English-centric. Miracl-vision (Osmulski et al., 2025) is the closest multilingual analogue extending visual retrieval to 18 languages including Arabic, but it explicitly documents up to 59.7% lower accuracy in the multilingual setting, exactly the gap a language-dedicated resource would address as shown in Table 1.

Arabic Document: Recognition vs. Retrieval.

A parallel Arabic research thread targets *transcription* rather than retrieval: KITAB-Bench (Heakl

Resource	Arabic dedicated	Visual doc. retrieval	Multi-source queries	Human validated	Hard-neg mining	Open release
<i>English / multilingual visual document retrieval</i>						
ViDoRe (Faysse et al., 2025)	✗	✓	✗	✓	✗	✓
M-BEIR (Wei et al., 2024)	✗	✓	✗	✓	✗	✓
IRPAPERS (Shorten et al., 2026)	✗	✓	✗	✓	✗	✓
VLM2Vec-V2 (Meng et al., 2025)	✗	✓	✗	✓	✗	✓
MIRACL-VISION (Osmulski et al., 2025)	✓	✓	✗	✓	✓	✓
<i>Arabic non-retrieval resources (OCR / instruction tuning)</i>						
KITAB-Bench (Heakl et al., 2025)	✓	✗	✗	✓	✗	✓
SynthDocs (Al-Homoud et al., 2025)	✓	✗	✗	✗	✗	✓
Baseer (Hennara et al., 2025)	✓	✗	✗	✓	✗	✓
QARI-OCR (Wasfy et al., 2025)	✓	✗	✗	✗	✗	✓
Peacock (Nagoudi et al., 2024)	✓	✗	✗	✓	✗	✓
PEARL (Alwajih et al., 2025)	✓	✗	✗	✓	✗	✓
DEAM (ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Arabic visual document retrieval resources. DEAM is the only Arabic-focused resource combining multi-source query generation, human-validated evaluation data, and mined hard negatives.

et al., 2025) benchmarks Arabic OCR across nine domains, whereas DEAM targets retrieval over rendered Arabic documents and evaluates whether models can match Arabic information needs to the relevant document. Cross-lingual SynthDocs (Al-Homoud et al., 2025) releases 2.5M synthetic OCR samples, and Baseer (Hennara et al., 2025) and QARI-OCR (Wasfy et al., 2025) reach 0.25 and 0.16 WER respectively through MLLM adaptation. Adjacent Arabic generation / instruction resources Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), Peacock (Nagoudi et al., 2024), PEARL (Alwajih et al., 2025) target instruction following, not retrieval. None ships an Arabic counterpart to ViDoRe; Table 1 positions DEAM against this landscape.

Synthetic Query Generation. Generating queries from documents is a standard approach for training dense retrievers. Doc2Query (Nogueira et al., 2019), InPars (Bonifacio et al., 2022), and Promptagator (Dai et al., 2022) use LLM prompting and filtering to generate queries from text. DEAM extends this approach to visual documents through multi-source generation (Claude Vision, Claude Text, and GPT-4o refinement) followed by human consensus validation.

3 DEAM Dataset

DEAM is constructed through a five-stage human-in-the-loop pipeline, shown in Figure 2. Human contributors participate at two points: in Stage I,

21 curate and prepare the source document corpus; in Stage IV, independent contributors review candidate queries for the held-out evaluation split in Argilla (Argilla, 2024), retaining items that receive two-of-three agreement. Document rendering, query generation, and hard-negative mining are automated.

3.1 Source Collection

Arabic source material is gathered from publicly available content across nine source domains. *Prose* domains include Wikipedia, news, books, science, and government documents; *visual* domains include tables, bar charts, line charts, and pie charts. After deduplication, content-quality filtering, and Unicode normalization, the corpus comprises 18,275 cleaned records as shown in Table 2. Query generation produces 82,795 flattened query candidates, which are consolidated into 82,727 final contrastive retrieval triplets.

3.2 Rendering

Documents are rendered to PNG using an HTML+Playwright² pipeline rather than Pillow. Pillow-based Arabic rendering produced detached or mis-shaped glyphs in our pilot, whereas browser rendering preserves standard Arabic contextual shaping and right-to-left layout. We use Naskh-style fonts, set direction: rtl, dynamically size each page to fit its content, and apply one of seven layout variants per document for visual diversity.

²<https://playwright.dev/>

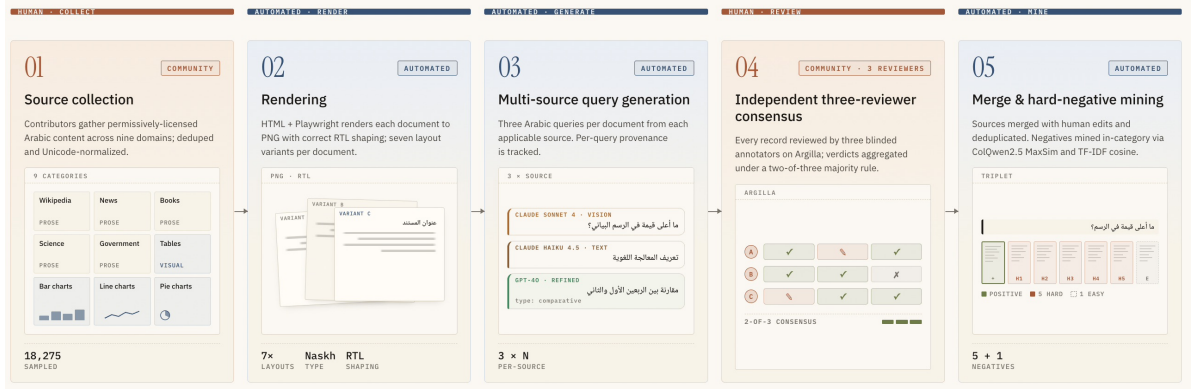


Figure 2: The five-stage construction pipeline behind DEAM: source collection, browser-based rendering, query generation, human validation, and hard-negative mining. The rendering stage preserves Arabic contextual shaping and right-to-left layout.

Category	Docs	Generated queries
Wikipedia	622	1,866
News	1,027	3,078
Books	389	1,167
Science	4,926	14,773
Government	1,944	5,832
Tables	2,834	16,986
Bar charts	696	4,216
Line charts	831	4,986
Pie charts	5,006	29,891
Total	18,275	82,795

Table 2: Per-category composition of DEAM. The query column reports flattened generated query candidates, before final triplet consolidation.

Charts and tables are rendered from structured numerical data with randomized chart styles. A 500-document spot check found no visible defects; remaining layout-quality failures are reviewed by human annotators as shown in Section 3.4.

3.3 Multi-Source Query Generation

Each document receives **three queries from every applicable source**; query provenance is tracked via a `sources` field on each query so downstream users can study per-source effects. Three independent LLM-based sources contribute queries: (i) **Claude Sonnet 4** (Anthropic, 2025) is shown the rendered image of each visual document and emits three diverse Arabic queries (one factual question, one keyword phrase, one descriptive paraphrase); (ii) **Claude Haiku 4.5** (Anthropic, 2025) consumes the underlying text content of prose documents under the same diverse-query schema including a cost-quality trade-off since purely textual prose does not benefit from vision input; and (iii) for the visual categories only, **GPT-4o** (OpenAI, 2024)

is additionally shown the rendered image and the three Claude queries. This process produces a refined, explicitly typed query set: *direct* queries ask about the document’s main subject; *comparative* queries require identifying a maximum, minimum, or trend; and *long-tail* queries target a specific entity, value, or number. In total, this stage adds 28,086 typed queries. Exact prompts are reproduced in Appendix A. After deduplication on a normalized form of the query text (NFKC, tashkeel stripped, punctuation collapsed), each surviving query carries its text, its source set, its type if applicable, and the identifier of its positive document.

3.4 Human Validation

All 21 contributors reviewed held-out queries on a self-hosted Argilla deployment (Argilla, 2024). Each query was independently assessed by three blinded annotators for image quality and query validity (*approve*, *edit*, or *reject*). Items with majority `bad_rendering` votes or no query-verdict majority were excluded; accepted edits were retained with a `human_edited` provenance flag. The release contains 2,799 human-validated test triplets, including a 2,660-query evaluable subset used for per-domain analysis. Agreement statistics and query-level annotation provenance are provided in Section 4 and the released data.

3.5 Merge, Deduplication, & Negative Mining

We merge all generated and human-edited queries, deduplicate on a normalized query string, and union the provenance flags of duplicates. For each remaining (q, d^+) pair, negatives are mined *within the same source category* to avoid trivial cross-domain distractors, combining two complementary signals. (i) **Visual**. ColQwen2.5-v0.2 (Faysse et al., 2024)

scores each query against in-category document images using MaxSim as shown:

$$\text{MaxSim}(q, d) = \sum_i \max_j \langle \bar{q}_i, \bar{d}_j \rangle, \quad (1)$$

with $\bar{v} = v/\|v\|_2$; we keep the top-30 non-positives. **(ii) Lexical.** TF-IDF cosine over each in-category document’s title + text_content; the top candidate adds a lexical hard negative, and one easy negative is sampled from ranks 20-30. Each released triplet ships with **five hard negatives** (top-4 ColQwen plus the top-1 non-duplicate TF-IDF candidate) and **one easy negative**, tagged hard / easy via neg_kinds. The triplets split embeds the positive image; triplets-eval (validation + test) additionally embeds all six negative images for direct side-by-side inspection.

4 Dataset Statistics and Quality Analysis

4.1 Annotation Quality

The validation analysis summarizes 4,209 judged query positions reviewed by 21 contributors under a two-of-three majority rule as shown in Table 3). Reviewer pairs achieve **75.2%** raw agreement. Chance-corrected agreement is lower (Cohen’s $\kappa = 0.184$; Fleiss’ $\kappa = 0.210$), which can occur when one label dominates the distribution (Feinstein and Cicchetti, 1990). At consensus, **89.5%** of positions are approved: 3,760 approved, 273 edited, and 169 rejected among the 4,202 positions that reached a majority; seven positions received no majority verdict. Per-category κ ranges from 0.060 for news, where *approve* dominates, to 0.325 for bar charts, where *edit* is more common.

A separate image_quality verdict removes **332 records** with majority bad_rendering, concentrated in line charts (45.5% of reviewed records) and science documents (32.5%). Figure 3 shows the per-category breakdown. The released records retain query-level source, verdict, edit-status, and image-quality provenance.

4.2 Query Length and Vocabulary Diversity

After human review, the final query set spans a range of formulations across sources. Token counts use whitespace tokenization after NFKC normalization and Arabic diacritic removal. Claude-only queries have a median length of 8 tokens. GPT-4o-refined queries are slightly longer, with a median of 10 tokens and $p_{90} = 12$, reflecting their explicit *direct*, *comparative*, and *long-tail* types. Human-edited

Metric	Value
Annotators	21
Judged query positions	4,209
Positions reaching majority	4,202
Positions with no majority	7
Pairwise verdict agreement (raw)	75.2%
Cohen’s κ (pairwise, macro)	0.184
Fleiss’ κ (multi-annotator)	0.210
Approve / Edit / Reject at consensus (%)	89.5 / 6.5 / 4.0
Approve / Edit / Reject at raw votes (%)	81.4 / 8.2 / 10.4
Records dropped (bad_rendering)	332

Table 3: Annotation quality summary.

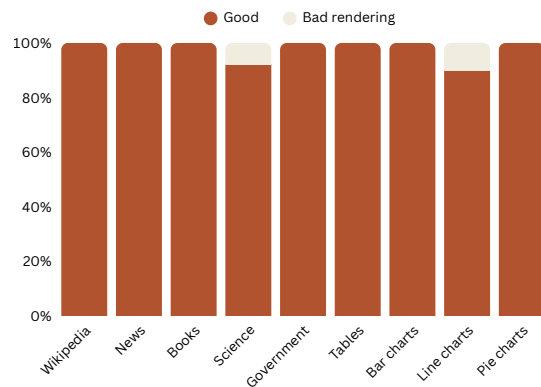


Figure 3: Distribution of the consensus image_quality verdict per category. Line charts and science have the highest bad_rendering rates.

queries are the longest, with $\bar{x} = 10.7$ and $p_{90} = 14$, as annotators generally added clarifying details rather than shortening queries.

Pairwise unigram Jaccard similarity across sources ranges from 0.35 to 0.46, indicating that no single style dominates the reviewed set. The highest overlap occurs in bar charts, where the Claude and human-edited queries have Jaccard similarity 0.65 and chart titles strongly constrain natural query wording.

5 Experimental Setup

We accompany DEAM with a zero-shot retrieval benchmark spanning **27 baselines**, comprising 17 text only retrievers and 10 visual document retrievers. Our benchmark compares clean-text retrieval, where models access document text directly, with image-only visual retrieval, where models must recover Arabic textual and layout information from rendered page images.

5.1 Test Split & Evaluation Protocol

The released held-out test split contains **2,799** human-validated triplets. The current reported

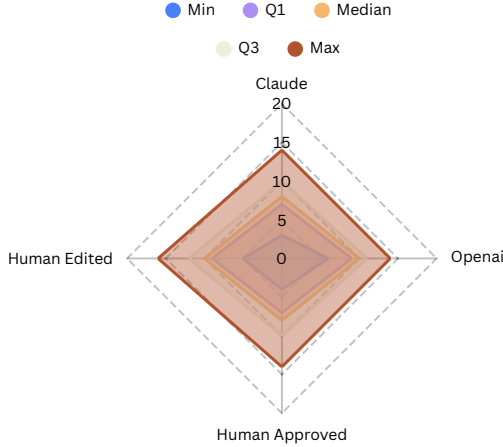


Figure 4: Query token-length distribution by source group. Whiskers cover 1.5 IQR. GPT-4o-refined queries are slightly longer than Claude originals, while human-edited queries are the longest.

retrieval benchmark uses the **2,660-query evaluable subset**, with one relevant document per query. For every evaluated query, each retriever ranks all **18,275 documents** in the corpus. We report $n\text{DCG}@10$ as the primary rank-sensitive metric, together with $\text{Recall}@ \{1, 5, 10\}$ and $\text{MRR}@10$. We compute 95% bootstrap confidence intervals using 1,000 resamples of the evaluation queries and random seed 42.

The evaluation subset covers five categories. News ($n = 1,094$), bar charts ($n = 809$), and tables ($n = 748$) are well represented, while books ($n = 3$) and government documents ($n = 6$) are small exploratory subsets. We therefore interpret aggregate scores alongside per-domain results and a macro-average over the three well-covered categories.

5.2 Baselines

We select baselines to span the contemporary retrieval design space along four axes: *modality* (text or visual), *language coverage* (English, multilingual, or Arabic-specific), *access* (open weights or closed API), and *architecture* (sparse lexical, dense single-vector, late-interaction multi-vector, or CLIP-style dual encoder). Table 4 lists the evaluated systems.

Implementation details. Text retrievers use the pooling strategy and query or document prefixes recommended by their respective model documentation. Visual retrievers follow their reference implementations to encode rendered PNG page images

Family	Evaluated baselines
Lexical	BM25 (Robertson and Zaragoza, 2009).
Open multilingual	BGE-M3 (Chen et al., 2024); mE5-large (Wang et al., 2022); GTE-Qwen2-7B (Li et al., 2023); Jina-v3 (Sturua et al., 2024); Nomic-v2-MoE (Nussbaum and Duderstadt, 2025); Qwen3-Emb-8B (Zhang et al., 2025); Llama-Nemotron-8B (Babakhin et al., 2025).
Arabic-specific	NeoAraBERT-MSA (Abou Chakra et al., 2026); GATE-AraBERT-v1 (Nacar et al., 2025); Arabic-Triplet-V2 (Nacar and Koubaa, 2025); SA-Retrieval-0.2B (Nacar, 2025); AraGemma-300M (Omartificial-Intelligence-Space, 2025); SILMA-Matryoshka (Soliman et al., 2024).
Commercial API	OpenAI text-embedding-3-large (Neelakantan et al., 2022); Cohere Embed-v4.0 (Cohere Team, 2025); Voyage-3-Large (Voyage AI, 2025).
Late-interaction VDR	ColPali-v1.3 (Faysse et al., 2025); ColQwen2-v1.0 (Faysse et al., 2024); ColQwen2.5-v0.2 (Faysse et al., 2024).
Single-vector VLM	DSE Qwen2-VL (Ma et al., 2024); GME-Qwen2-VL-2B (Zhang et al., 2024).
CLIP-style VLM	Jina-CLIP-v2 (Koukounas et al., 2024); SigLIP2 (Zhai et al., 2023).
Other visual	BGE-Visualized-M3 (Zhou et al., 2024); VisRAG-MiniCPM-V (Yu et al., 2025); ColSmoI-VLM-alpha (Faysse et al., 2025).

Table 4: Zero-shot baselines evaluated on DEAM.

at native resolution. Commercial endpoints use each provider’s retrieval configuration, and corpus embeddings are cached. Exact encoder revisions, prompts, prefixes, pooling methods, preprocessing, hardware, precision, runtimes, commands, and random seeds are documented in Appendix B.

5.3 Fine-Tuning Protocol

We contrastively fine-tune BGE-M3 (BAAI/bge-m3) on DEAM training triplets to assess its value beyond zero-shot evaluation. Each example contains an Arabic query, a positive document represented by its concatenated title and text, and four negatives; no query or document prefixes are used. Our primary condition uses released mined hard negatives, while the matched control uses four random documents from the same source domain. Missing hard negatives are filled with random in-domain negatives, ensuring that both conditions use the same number of negatives and training recipe.

We train for one epoch with Multiple Negatives Ranking Loss and in-batch negatives, using batch size 24, learning rate 2×10^{-5} , maximum sequence length 512, 10% linear warm-up, mixed precision, and seed 42. Models are evaluated on the same 2,660-query subset and with the same metrics as the zero-shot experiments. The released training script stores the complete configuration with each checkpoint.

5.4 LLM Reranking Protocol

We additionally evaluate generative reranking as an upper bound over a fixed candidate set. We first retrieve the top ten documents from the full DEAM corpus using zero-shot BGE-M3 (BAAI/bge-m3). We then rerank these candidates pointwise with qwen3.6:latest, served locally through an OpenAI-compatible Ollama endpoint.

For each Arabic query-document pair, the model receives the Arabic query and up to the first 1,200 characters of the document title and text. It assigns an integer relevance score from 0 (irrelevant) to 3 (perfect match) and returns the score as JSON. We use temperature 0 and retain the original BGE-M3 rank to break ties. The experiment uses random seed 42 and a proportionally stratified sample of 150 queries drawn from the evaluable test subset.

We use Qwen rather than Claude or GPT-4o because the latter models contributed to query generation during dataset construction. Evaluating them on queries that they helped generate would introduce leakage and produce an optimistically biased comparison. This experiment, therefore, measures leakage-free generative reranking on a fixed candidate set; it is not a standalone corpus-retrieval baseline.

5.5 English-Pivot Retrieval Protocol

We evaluate whether translating both sides of the retrieval problem into English improves Arabic document retrieval. We translate Arabic queries and the Arabic document corpus using facebook/nllb-200-distilled 600M, with arb_Arab as the source language and eng_Latn as the target language. Translation uses batches of 32 and a maximum input and generation length of 256 tokens. To make corpus translation tractable while controlling the comparison, we retain the first 400 characters of each document’s concatenated title and text in both conditions.

We evaluate a proportionally stratified sample of 200 queries from the evaluable test subset, selected with random seed 42. The native condition retrieves Arabic documents with BGE-M3 (BAAI/bge-m3). The English pivot condition retrieves NLLB-translated documents with BAAI/bge-large-en-v1.5. Neither condition uses a query prefix. We use normalized dense embeddings and evaluate both systems against the same Arabic relevance judgments.

Model	nDCG@10	R@1	R@10	MRR@10
<i>Lexical baseline</i>				
BM25 (Robertson and Zaragoza, 2009)	0.617	0.541	0.697	0.592
<i>Open multilingual</i>				
BGE-M3 (Chen et al., 2024)	0.780	0.697	0.865	0.753
mE5-large (Wang et al., 2022)	0.804	0.732	0.878	0.780
GTE-Qwen2-7B (Li et al., 2023)	0.798	0.717	0.880	0.771
Jina-v3 (Sturua et al., 2024)	0.806	0.728	0.883	0.781
Nomic-v2-MoE (Nussbaum and Duderstadt, 2025)	0.792	0.708	0.879	0.764
Qwen3-Emb-8B (Zhang et al., 2025)	0.831	0.755	0.909	0.806
Llama-Nemotron-8B (Babakhin et al., 2025)	0.223	0.156	0.299	0.199
<i>Arabic-specific</i>				
NeoAraBERT-MSA (Abou Chakra et al., 2026)	0.004	0.001	0.008	0.002
GATE-AraBERT-v1 (Nacar et al., 2025)	0.653	0.565	0.744	0.624
Arabic-Triplet-V2 (Nacar and Koubaa, 2025)	0.651	0.545	0.766	0.614
SA-Retrieval-0.2B (Nacar, 2025)	0.390	0.314	0.474	0.363
AraGemma-300M (Omarificial-Intelligence-Space, 2025)	0.718	0.603	0.840	0.679
SILMA-Matryoshka (Soliman et al., 2024)	0.660	0.549	0.775	0.624
<i>Commercial API</i>				
OpenAI embed-3-large (Neelakantan et al., 2022)	0.742	0.633	0.859	0.705
Cohere Embed-v4.0 (Cohere Team, 2025)	0.828	0.757	0.900	0.805
Voyage-3-Large (Voyage AI, 2025)	0.833	0.753	0.912	0.807

Table 5: Zero-shot clean-text retrieval results on the 2,660-query evaluation subset of DEAM.

6 Results & Discussion

6.1 Main Retrieval Results

Tables 5 and 7 report zero-shot performance for all 27 baselines on the 2,660 evaluation subset of the human-validated test split. The main finding is that **clean text retrievers substantially outperform image-only visual retrievers on Arabic documents**. The best text system, Voyage-3-Large, achieves nDCG@10 of 0.833, while the best visual system, ColQwen2.5-v0.2, achieves 0.697, a difference of 0.136 nDCG@10.

Why clean Arabic text outperforms visual retrieval? This comparison is practically meaningful but not a fully controlled modality comparison. Text retrievers receive clean Unicode title and text fields, whereas visual retrievers must recover the same information from page pixels. Arabic introduces additional difficulty because correct interpretation requires reading contextually shaped glyphs and right-to-left layout. Late-interaction visual retrievers with Qwen2.5-VL backbones close part of this gap, but do not match the strongest multilingual text encoders on the aggregate evaluation.

Model	News $n = 1,094$	Bar charts $n = 809$	Tables $n = 748$	Books $n = 3$	Government $n = 6$
Voyage-3-Large	0.889	0.715	0.881	0.877	0.439
Qwen3-Embedding-8B	0.863	0.748	0.879	0.767	0.500
Cohere Embed-v4.0	0.862	0.754	0.863	0.667	0.572
BGE-M3	0.828	0.666	0.838	0.477	0.500
ColQwen2.5-v0.2	0.959	0.275	0.770	0.877	0.692
ColQwen2-v1.0	0.861	0.352	0.671	0.767	0.680
BM25	0.751	0.378	0.682	0.500	

Table 6: Per-domain nDCG@10 across five represented test categories. Results for books ($n = 3$) and government documents ($n = 6$) should be interpreted cautiously because of their small evaluation subsets.

Model	nDCG@10	R@1	R@10	MRR@10
<i>Late-interaction multi-vector</i>				
ColPali-v1.3 (Faysse et al., 2025)	0.227	0.148	0.323	0.198
ColQwen2-v1.0 (Faysse et al., 2024)	0.652	0.554	0.758	0.619
ColQwen2.5-v0.2 (Faysse et al., 2024)	0.697	0.612	0.787	0.669
<i>Single-vector VLM</i>				
DSE Qwen2-VL (Ma et al., 2024)	0.003	0.000	0.008	0.002
GME-Qwen2-VL-2B (Zhang et al., 2024)	0.522	0.377	0.680	0.472
<i>CLIP-style multilingual VLM</i>				
Jina-CLIP-v2 (Koukounas et al., 2024)	0.038	0.021	0.059	0.031
SigLIP2 (Zhai et al., 2023)	0.096	0.059	0.147	0.081
<i>Other visual retrievers</i>				
BGE-Visualized-M3 (Zhou et al., 2024)	0.017	0.006	0.032	0.013
VisRAG-MiniCPM-V (Yu et al., 2025)	0.017	0.011	0.024	0.014
ColSmolVLM-alpha (Faysse et al., 2025)	0.029	0.020	0.040	0.025

Table 7: Zero-shot image-only visual document retrieval results on the 2,660-query evaluation subset of DEAM. **Bold** marks the best visual retriever.

6.2 Per-Domain Performance

The evaluable test subset contains 2,660 queries across five categories. News ($n = 1,094$), bar charts ($n = 809$), and tables ($n = 748$) support category-level analysis, whereas books ($n = 3$) and government documents ($n = 6$) remain exploratory. Table 6 reports per-domain nDCG@10 for all retrievers.

Voyage-3-Large is the strongest clean-text retriever overall, leading on tables (0.881) and remaining competitive on news (0.889). ColQwen2.5 achieves the best news score (0.959), but drops sharply on bar charts (0.275) and tables (0.770). Macro-averaging news, bar charts, and tables gives 0.828 for Voyage-3-Large and 0.668 for ColQwen2.5, exceeding their micro-average gap (0.160 versus 0.136). Thus, the clean text advantage is not driven solely by the largest category. Complete results and 95% bootstrap confidence intervals are included in the released files.

Model	nDCG@10	R@1	MRR@10
BGE-M3 (zero-shot)	0.780	0.697	0.753
BGE-M3 + random in-domain negatives	0.826	0.755	0.803
BGE-M3 + mined hard negatives	0.833	0.764	0.811

Table 8: Effect of fine-tuning BGE-M3 on DEAM. Hard-negative mining provides a further improvement over random in-domain negatives.

6.3 Fine-Tuning with DEAM Triplets

We next test whether DEAM provides useful supervision for Arabic retrieval training. Table 8 compares zero-shot BGE-M3 with models contrastively fine-tuned on the DEAM training triplets using random in-domain negatives or mined hard negatives.

Fine-tuning improves BGE-M3 from 0.780 to 0.833 nDCG@10, an absolute gain of 0.053 over the zero-shot model (one-sided paired bootstrap test, $p < 0.001$). This shows that DEAM is useful as training data as well as an evaluation resource.

The hard-negative comparison isolates the value of the mined negatives. Under the same base model, seed, epochs, batch size, learning rate, and number of negatives, mined hard negatives improve nDCG@10 by 0.007 over random in domain negatives. This difference is statistically significant (95% CI [0.0004, 0.014]; one-sided paired bootstrap test, $p = 0.019$). The hard-negative model also obtains the highest R@1 and MRR@10, indicating that the mined negatives provide more informative supervision than matched random negatives.

6.4 LLM Reranking

We assess whether a generative model can improve ranking after a dense first-stage retriever identifies a small candidate set. On a proportionally stratified sample of 150 queries, Qwen reranks the top ten documents retrieved by BGE-M3 using pointwise Arabic relevance judgments. As shown in Table 9,

Method	nDCG@10	R@1
BGE-M3 first-stage ranking	0.789	0.700
BGE-M3 + Qwen reranking	0.815	0.747

Table 9: Leakage-free LLM reranking on 150 proportionally stratified evaluation queries. Qwen reranks the top ten candidates retrieved by BGE-M3.

Model	Arabic queries	English queries	Δ
Voyage-3-Large	0.808	0.766	-0.042
ColQwen2.5-v0.2	0.638	0.614	-0.024
ColQwen2-v1.0	0.651	0.462	-0.189

Table 10: Cross-lingual nDCG@10 on the same stratified sample of 200 evaluation queries. Δ denotes the performance change from Arabic to English queries.

reranking improves nDCG@10 from 0.789 to 0.815 and R@1 from 0.700 to 0.747. This demonstrates that generative relevance judgments can improve ranking quality when paired with a strong Arabic dense retriever, although performance remains bounded by the recall of BGE-M3’s top-ten candidates; it is not a full-corpus LLM retrieval baseline.

We use Qwen rather than Claude or GPT-4o because Claude and GPT-4o contributed to query generation during dataset construction. Evaluating these models on queries they helped generate could introduce leakage and yield an optimistically biased comparison.

6.5 Cross-Lingual & English-Pivot Retrieval

A possible shortcut for Arabic retrieval is to translate Arabic queries into English while retaining the Arabic corpus. Table 10 tests this setting on 200 stratified evaluation queries. Native Arabic queries consistently outperform NLLB-translated English queries, especially for visual retrievers.

We also evaluate the full English-pivot setting proposed by the reviewers: both the Arabic queries and Arabic documents are translated into English using NLLB-200, then retrieved with an English BGE model. Table 11 uses the same 200-query sample, 400-character document representation, and Arabic relevance judgments in both conditions.

Translating both sides of the retrieval problem reduces nDCG@10 from 0.780 to 0.497, an absolute decrease of 0.284. The same pattern holds for R@1, R@10, and MRR@10. Translation, therefore, does not recover the retrieval signal lost from Arabic text and document structure; direct Arabic retrieval

Condition	nDCG@10	R@1	R@10	MRR@10
Native Arabic, BGE-M3	0.780	0.705	0.865	0.754
English pivot, BGE-large-en-v1.5	0.497	0.395	0.615	0.460

Table 11: Native Arabic retrieval versus English-pivot retrieval on the same stratified sample of 200 queries.

is markedly more effective than an English-pivot approach.

6.6 Discussion

Voyage-3-Large, Qwen3-Embedding-8B, and Cohere Embed-v4.0 form a closely matched leading group, within 0.005 nDCG@10 of one another. The strongest open model, Qwen3-Embedding-8B, approaches the commercial APIs, whereas Arabic-specific text embedders score between 0.65 and 0.72 nDCG@10. This gap suggests that retrieval-oriented training on DEAM could further strengthen Arabic-focused encoders.

The released mined negatives are consistently more difficult than random in-domain negatives across all 12 tested retrievers, with significant differences for 10 models ($p < 0.001$, one-sided Wilcoxon signed-rank test). Among the 50 lowest-scoring ColQwen2.5 queries, 94% of failures involve top- K confusion with hard negatives from the correct domain. Together, the results show that clean-text multilingual retrievers currently lead aggregate Arabic retrieval, while visual retrieval is particularly effective for layout-rich news pages. The corpus, triplets, and reproducible evaluation harness provide a foundation for Arabic-aware visual retrievers with stronger glyph and right-to-left layout understanding.

7 Conclusion

We presented DEAM, the first open Arabic visual document retrieval resource, with 18,275 rendered Arabic documents spanning nine source domains and 82,727 contrastive retrieval triplets. Across 27 zero-shot retrievers, strong multilingual text models outperform the best visual retrievers by 13.6 nDCG@10 points overall, although visual retrieval is strongest on layout-rich news pages. Fine-tuning BGE-M3 on DEAM improves nDCG@10 from 0.780 to 0.833, with mined hard negatives outperforming matched random negatives. By releasing the data, splits, and reproducible evaluation and training harnesses, DEAM establishes a foundation for Arabic visual document retrieval research.

8 Limitations

Our fine-tuning study covers only BGE-M3; we do not fine-tune visual retrievers or systematically compare architectures and hyperparameters. DEAM uses browser-rendered Modern Standard Arabic documents rather than natural scans, so it does not capture all real-world noise, handwriting, dialectal variation, or regional document styles. The current evaluation subset covers five categories and is concentrated in news, bar charts, and tables, while books and government remain exploratory subsets too small for strong category-level conclusions. Queries are LLM-generated and then human reviewed, but generation choices and annotator judgments may still introduce stylistic or selection bias. Finally, source-level licensing metadata are released, but downstream users must ensure that their intended use complies with the relevant source licenses.

Acknowledgments

We gratefully acknowledge Tuwaiq Academy for its valuable support of this work, including research resources, computational infrastructure, and financial support.

References

- Chadi Abou Chakra, Hadi Hamoud, Osama Rakan Al Mraikhat, Qusai Abu Obaida, Mohamad Ballout, and Fadi A Zaraket. 2026. Neoarabert: A modern foundation model for arabic embeddings with diacritics-aware tokenization and pos-targeted masking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 25952–25968.
- Hassan Al-Homoud, Ahmed A. Ibrahim, Mohammed Al-Jubran, Faisal Al-Otaibi, Yousef Alharbi, Daurén Toibazar, Kai Wang, and Pedro J. Moreno. 2025. [Cross-lingual SynthDocs: A large-scale synthetic corpus for any-to-Arabic ocr and document understanding](#). In *International Conference on Data and Software Engineering*.
- Fakhraddin Alwajih, Samar M. Magdy, Abdellah El Mekki, and 1 others. 2025. [Pearl: A multimodal culturally-aware Arabic instruction dataset](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23048–23079.
- Anthropic. 2025. [Claude haiku 4.5: Model overview](#).
- Anthropic. 2025. [System card: Claude opus 4 & claude sonnet 4](#). Technical report, Anthropic. Accessed: August 29, 2026.
- Argilla. 2024. Argilla: The open-source data collaboration platform for ML. <https://argilla.io/>.
- Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-Embed-Nemotron-8B: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*.
- Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).
- Cohere Team. 2025. Introducing Embed 4: Multimodal search for business. <https://cohere.com/blog/embed-4>. Cohere Blog.
- Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColQwen2.5-v0.2: Late-interaction visual document retriever based on Qwen2.5-VL. <https://huggingface.co/vidore/colqwen2.5-v0.2>. Model card.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.
- Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. 2025. Kitab-bench: A comprehensive multi-domain benchmark for arabic ocr and document understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22006–22024.
- Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Baseer: A vision-language model for arabic document-to-markdown ocr. *arXiv preprint arXiv:2509.18174*.
- Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, and 1 others. 2024. Acegpt, localizing large language models in arabic.

- In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.
- Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Andreas Koukounas, Georgios Mastrapas, Sedigheh Eslami, Bo Wang, Mohammad Kalim Akram, Michael Günther, Isabelle Mohr, Saba Sturua, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025. Gemini Embedding: Generalizable embeddings from Gemini. *arXiv preprint arXiv:2503.07891*.
- Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.
- Rui Meng, Ziyang Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and 1 others. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.
- Omer Nacar. 2025. SA-Retrieval-Embeddings-0.2b: Retrieval-optimized Saudi Arabic sentence embeddings.
- Omer Nacar and Anis Koubaa. 2025. Enhancing semantic similarity understanding in Arabic NLP with nested embedding learning. In *Generative AI and Large Language Models: Opportunities, Challenges, and Applications: Volume 1*, pages 179–216. Springer.
- Omer Nacar, Anis Koubaa, Serry Sibae, Yasser Al-Habashi, Adel Ammar, and Wadii Boulila. 2025. GATE: General Arabic text embedding for enhanced semantic textual similarity with matryoshka representation learning and hybrid loss training. *arXiv preprint arXiv:2505.24581*.
- Fakhraddin Alwajih El Moatez Billah Nagoudi, Gagan Bhatia, and Abdelrahman Mohamed Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.
- Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.
- Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *arXiv preprint arXiv:2502.07972*.
- Omar Artificial-Intelligence-Space. 2025. [AraGemma-Embedding: Fine-tuned EmbeddingGemma for Arabic semantic understanding](#).
- OpenAI. 2024. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>.
- Radek Osmulski, Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. 2025. Miracl-vision: A large, multilingual, visual document retrieval benchmark. *arXiv preprint arXiv:2505.11651*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Connor Shorten, Augustas Skaburskas, Daniel M Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilloker, and Bob van Luijt. 2026. Irpayers: A visual document benchmark for scientific retrieval and question answering. *arXiv preprint arXiv:2602.17687*.

Abu Bakr Soliman, Karim Ouda, and SILMA AI. 2024. SILMA embedding matryoshka 0.1. <https://huggingface.co/silma-ai/silma-embeddding-matryoshka-0.1>.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and 1 others. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint arXiv:2409.10173*.

Voyage AI. 2025. voyage-3-large: The new state-of-the-art general-purpose embedding model. <https://www.voyageai.com/blog/voyage-3-large>. Voyage AI Blog.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Ahmed Wasfy, Omer Nacar, Aly Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. QARI-OCR: High-fidelity Arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations (ICLR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952. IEEE.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Yanzhao Zhang, Dingkun Long, Mingxin Li, Pengjun Xie, and 1 others. 2025. Qwen3-Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200.

A Query Generation Prompts

This appendix reproduces the prompts used by the three LLM sources in stage iii of the construction pipeline (Section 3.3). Prompts are presented in English for readability. Figures 5 and 6 show the query-generation prompts used for visual and prose documents, respectively.

The GPT-4o refinement prompt is shown in Figure 7. In the Claude prompts, {n} is set to 3, producing three queries per document. In the Haiku prompt, {text} contains the first 800 characters of the document text_content. The generation scripts, prompt templates, and query-level provenance are included in the public release.

Claude Sonnet 4 vision prompt

Task. Generate Arabic search queries for a visual document. Given the rendered document image, produce {n} natural Arabic queries that a user might type to retrieve this document.

Rules. Use Modern Standard Arabic. The queries should sound like real search-engine queries and should not be paraphrases of one another. Produce diverse query forms: a direct question, a keyword-style query, and a descriptive query. If the document contains a chart or table, mention that in at least one query.

Output. Return only valid JSON: {"queries": ["query 1", "query 2", "query 3"]}

Figure 5: Claude Sonnet 4 vision prompt used for chart and table categories. The model receives the rendered page image and returns three Arabic search queries in JSON format.

Claude Haiku 4.5 text prompt

Task. Generate Arabic search queries for a prose document. Given the document text, produce {n} natural Arabic queries that a user might type to retrieve this document.

Input. Text: {text}

Rules. Use Modern Standard Arabic. The queries should sound like real search-engine queries and should not be paraphrases of one another. Produce diverse query forms: a direct question, a keyword-style query, and a descriptive query.

Output. Return only valid JSON: {"queries": ["query 1", "query 2", "query 3"]}

Figure 6: Claude Haiku 4.5 text prompt used for prose categories: Wikipedia, news, books, science, and government. The model receives document text because rendered prose pages do not add information beyond the text content for query generation.

GPT-4o typed refinement prompt

System message. You design Arabic search queries for a multimodal retrieval dataset. For each document image, produce *exactly three* Arabic queries, each with a different type: direct, comparative, and long_tail.

Query types. direct: a natural question or keyword phrase about the main subject. comparative: a query comparing values, trends, highest or lowest items, or visible relationships in a chart or table. long_tail: a query targeting a specific entity, year, number, label, or sub-detail.

Rules. Return only valid JSON. All queries must be in Modern Standard Arabic and sound like real search-engine queries. Do not invent values, names, years, or categories; every referenced detail must be visible in the image or unambiguous from the text. The three queries must be substantively different, not paraphrases. Use the original three candidate queries only as inspiration. Each query must be 15–120 characters.

User message. Document type: {doc_type}
Original candidate queries: {q1}; {q2}; {q3}
Look at the document image and produce the refined trio.

Output schema. A JSON object with field queries, containing exactly three objects. Each object has type, query, and rationale.

Figure 7: GPT-4o typed refinement prompt used for the four visual categories. The model receives the rendered image, document type, and candidate queries, then returns a direct, comparative, and long-tail Arabic query under a strict JSON schema.

B Baseline Protocols and Detailed Results

This appendix documents the implementation settings for the zero-shot baselines and provides the complete set of retrieval metrics omitted from the main-paper tables for space reasons. All reported results use the 2,660-query evaluable subset and rank against the full 18,275-document corpus.

Text retrievers. Open dense text retrievers use the pooling method and query or document prefixes recommended by their model documentation. BGE-M3 uses no prefix; mE5-large uses query: and passage:; GTE-Qwen2-7B uses its retrieval instruction template; Qwen3-Embedding-8B uses its retrieval task instruction; Jina-v3 uses the retrieval.query and retrieval.passage task labels; and Nomic-Embed-v2-MoE uses search_query: and search_document:. Arabic-specific encoders follow their published preprocessing procedures. Commercial APIs use the providers’ documented retrieval settings, and corpus embeddings are cached.

BM25 is evaluated over the same title and text representation used by the dense text retrievers. Dense retrievers use cosine similarity over L2-normalized vectors and exhaustively rank the full corpus. Metrics use 1,000 bootstrap resamples with random seed 42. Exact model identifiers, encoder revisions, prompts, preprocessing settings, and reproduction commands are included in the public release. Table 12 gives the complete clean-

text results.

Visual retrievers. ColPali-v1.3, ColQwen2-v1.0, and ColQwen2.5-v0.2 are evaluated with their reference implementations and late-interaction MaxSim scoring. GME-Qwen2-VL-2B and DSE Qwen2-VL-2B use single-vector document representations. Jina-CLIP-v2 and SigLIP2 use separate image and text encoders with cosine similarity. BGE-Visualized-M3, VisRAG, and ColSmoVLm-alpha follow their respective reference codebases. All visual methods receive rendered PNG page images at native resolution. Table 13 gives the corresponding visual results.

Hardware and runtime. Visual encoders are evaluated on one NVIDIA RTX 5090 GPU with 32 GB memory, using fp16 or bf16 where supported. Encoding the 18,275-document visual corpus requires approximately 30 to 60 minutes per visual encoder. Text encoders run on the same hardware, except for commercial APIs; BM25 indexing runs on CPU in minutes.

C Representative Per-Domain Results

Table 6 in the main paper reports representative per-domain results across all five represented categories. This appendix adds 95% percentile bootstrap confidence intervals for the three well-covered categories. Confidence intervals use 1,000 query-set resamples and random seed 42.

The evaluation subset contains 2,660 queries from five categories: news ($n = 1,094$), bar charts ($n = 809$), tables ($n = 748$), books ($n = 3$), and government documents ($n = 6$). The books and government subsets are included for transparency, but their small sample sizes yield wide confidence intervals. Table 14 reports the well-covered categories, followed by the exploratory subsets in Table 15.

D Mining Quality: Cross-Model Validation

This appendix expands the mining-quality analysis summarized in Section 6.6. For each of the 2,660 evaluation queries, we construct a candidate pool containing the gold positive document, five mined hard negatives, and six random in-domain negatives sampled with random seed 42. The mined negatives originate from the released neg_ids field and were initially identified with ColQwen2.5.

For each auxiliary retriever, we compare the per-query mean ranks of mined and random negatives

Model	nDCG@10	R@1	R@5	R@10	MRR@10
BM25 (Robertson and Zaragoza, 2009)	0.617	0.541	0.656	0.697	0.592
BGE-M3 (Chen et al., 2024)	0.780	0.697	0.827	0.865	0.753
mE5-large (Wang et al., 2022)	0.804	0.732	0.842	0.878	0.780
GTE-Qwen2-7B (Li et al., 2023)	0.798	0.717	0.844	0.880	0.771
Jina-v3 (Sturua et al., 2024)	0.806	0.728	0.852	0.883	0.781
Nomic-v2-MoE (Nussbaum and Duderstadt, 2025)	0.792	0.708	0.839	0.879	0.764
Qwen3-Embedding-8B (Zhang et al., 2025)	0.831	0.755	0.875	0.909	0.806
Llama-Nemotron-8B (Babakhin et al., 2025)	0.223	0.156	0.252	0.299	0.199
NeoAraBERT-MSA (Abou Chakra et al., 2026)	0.004	0.001	0.005	0.008	0.002
GATE-AraBERT-v1 (Nacar et al., 2025)	0.653	0.565	0.698	0.744	0.624
Arabic-Triplet-V2 (Nacar and Koubaa, 2025)	0.651	0.545	0.703	0.766	0.614
SA-Retrieval-0.2B (Nacar, 2025)	0.390	0.314	0.428	0.474	0.363
AraGemma-300M (Omarificial-Intelligence-Space, 2025)	0.718	0.603	0.781	0.840	0.679
SILMA-Matryoshka (Soliman et al., 2024)	0.660	0.549	0.719	0.775	0.624
OpenAI text-embedding-3-large (Neelakantan et al., 2022)	0.742	0.633	0.802	0.859	0.705
Cohere Embed-v4.0 (Cohere Team, 2025)	0.828	0.757	0.870	0.900	0.805
Voyage-3-Large (Voyage AI, 2025)	0.833	0.753	0.879	0.912	0.807

Table 12: Detailed zero-shot metrics for all clean-text retrievers on the 2,660-query DEAM evaluation subset.

Model	nDCG@10	R@1	R@5	R@10	MRR@10
ColPali-v1.3 (Faysse et al., 2025)	0.227	0.148	0.265	0.323	0.198
ColQwen2-v1.0 (Faysse et al., 2024)	0.652	0.554	0.704	0.758	0.619
ColQwen2.5-v0.2 (Faysse et al., 2024)	0.697	0.612	0.738	0.787	0.669
DSE Qwen2-VL (Ma et al., 2024)	0.003	0.000	0.005	0.008	0.002
GME-Qwen2-VL-2B (Zhang et al., 2024)	0.522	0.377	0.597	0.680	0.472
Jina-CLIP-v2 (Koukounas et al., 2024)	0.038	0.021	0.041	0.059	0.031
SigLIP2 (Zhai et al., 2023)	0.096	0.059	0.108	0.147	0.081
BGE-Visualized-M3 (Zhou et al., 2024)	0.017	0.006	0.023	0.032	0.013
VisRAG-MiniCPM-V (Yu et al., 2025)	0.017	0.011	0.020	0.024	0.014
ColSmolVLM-alpha (Faysse et al., 2025)	0.029	0.020	0.031	0.040	0.025

Table 13: Detailed zero-shot metrics for all image-only visual retrievers on the 2,660-query DEAM evaluation subset.

Model	News	Bar charts	Tables
Voyage-3-Large	0.889 [0.872, 0.904]	0.715 [0.690, 0.742]	0.881 [0.863, 0.899]
Qwen3-Embedding-8B	0.863 [0.845, 0.880]	0.748 [0.723, 0.774]	0.879 [0.860, 0.897]
Cohere Embed-v4.0	0.862 [0.844, 0.878]	0.754 [0.728, 0.781]	0.863 [0.843, 0.882]
BGE-M3	0.828 [0.807, 0.847]	0.666 [0.636, 0.694]	0.838 [0.818, 0.860]
ColQwen2.5-v0.2	0.959 [0.950, 0.967]	0.275 [0.248, 0.301]	0.770 [0.746, 0.795]
ColQwen2-v1.0	0.861 [0.843, 0.878]	0.352 [0.323, 0.382]	0.671 [0.643, 0.700]
BM25	0.751 [0.726, 0.775]	0.378 [0.349, 0.410]	0.682 [0.653, 0.713]

Table 14: Per-domain nDCG@10 with 95% bootstrap confidence intervals for news, bar charts, and tables.

Model	Books ($n = 3$)	Government ($n = 6$)
Voyage-3-Large	0.877 [0.631, 1.000]	0.439 [0.105, 0.833]
Qwen3-Embedding-8B	0.767 [0.301, 1.000]	0.500 [0.167, 0.833]
Cohere Embed-v4.0	0.667 [0.000, 1.000]	0.572 [0.238, 0.905]
BGE-M3	0.477 [0.000, 1.000]	0.500 [0.167, 0.833]
ColQwen2.5-v0.2	0.877 [0.631, 1.000]	0.692 [0.455, 0.905]
ColQwen2-v1.0	0.767 [0.301, 1.000]	0.680 [0.452, 0.905]
BM25	0.667 [0.000, 1.000]	0.500 [0.167, 0.833]

Table 15: Per-domain nDCG@10 with 95% bootstrap confidence intervals for exploratory book and government-document subsets.

using a one-sided Wilcoxon signed-rank test (H_1 : mined < random). A positive difference, computed as random rank minus mined rank, indicates that mined negatives are harder. Table 16 reports the cross-model test results, while Figures 8, 9, and 10 provide per-query, top- K , and distributional views of the same finding.

E Cross-Lingual Generalization

Figure 11 visualizes the query-translation result reported in Table 10. Native Arabic queries outperform NLLB-translated English queries for all three retrievers. The effect is small for ColQwen2.5 but substantially larger for ColQwen2.

Model	Mined rank	Random rank	Δ	Wilcoxon p
ColQwen2.5-v0.2 (<i>mining model</i>)	4.13	7.36	3.23	<0.001
BGE-Multilingual-Gemma2	4.63	7.05	2.42	<0.001
NV-Embed-v2	4.81	6.86	2.05	<0.001
ColQwen2-v1.0	4.90	6.70	1.81	<0.001
GME-Qwen2-VL-2B	5.01	6.56	1.54	<0.001
ColPali-v1.3	5.41	5.93	0.51	<0.001
SigLIP2-so400m	5.43	5.71	0.28	<0.001
Jina-CLIP-v2	5.43	5.63	0.20	<0.001
mxbai-embed-large-v1	5.51	5.70	0.18	<0.001
BGE-Visualized-M3	5.44	5.59	0.16	<0.001
VisRAG-Ret-MiniCPM-V	5.46	5.57	0.11	0.007
DSE Qwen2-VL	5.47	5.55	0.08	0.033

Table 16: Cross-model validation of mined negatives. Positive Δ indicates that mined negatives receive lower ranks and are therefore harder than random in-domain negatives. Ten of the twelve retrievers have $p < 0.001$.

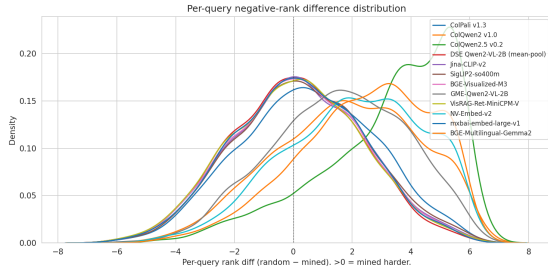


Figure 8: Per-query rank difference between random and mined negatives across the 12 auxiliary retrievers. Positive values indicate that mined negatives are harder.

F Qualitative Failure-Mode Analysis

This appendix expands the qualitative error analysis summarized in Section 6.6. We manually inspect the 50 lowest-scoring queries for ColQwen2.5 and ColQwen2, the two strongest visual retrievers, and assign each case to a six-class failure taxonomy. Table 17 reports the taxonomy, and Figure 12 visualizes its distribution.

Failure mode	ColQwen2.5	ColQwen2
Hard-negative top- K	47 (94%)	18 (35%)
Domain confusion	2 (4%)	24 (48%)
Chart misread	0	7 (14%)
OCR or text issue	0	1 (2%)
Data quality issue	1 (2%)	0
Other	0	1 (2%)
Total	50	51

Table 17: Failure-mode counts among the lowest-scoring retrieval cases.

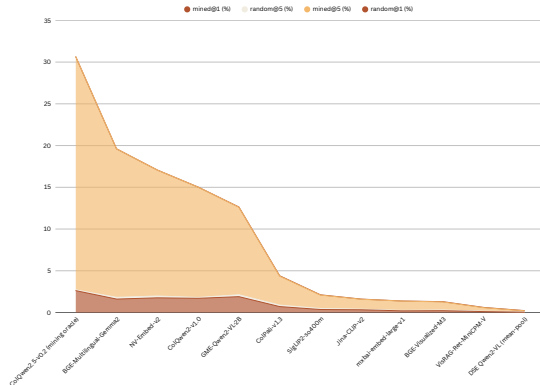


Figure 9: Fraction of mined hard negatives and random in-domain negatives appearing in each retriever's full-corpus top- K .

Failure-mode definitions. *Hard-negative top- K* denotes a same-domain and same-topic distractor that outranks the gold positive. *Domain confusion* denotes retrieval from an incorrect source domain. *Chart misread* refers to matching visual layout while missing a metric or label. *OCR or text issue* denotes insufficiently readable rendered text. *Data quality issue* identifies an empty, mislabeled, or weakly annotated gold document. *Other* contains residual cases not covered by these categories.

Example retrieval cases. Figures 13, 14, and 15 show three representative ColQwen2.5 failures. The common pattern is visual and topical similarity between the gold page and retrieved pages, with an error in a numeric, temporal, geographic, or named-entity detail. These cases motivate the

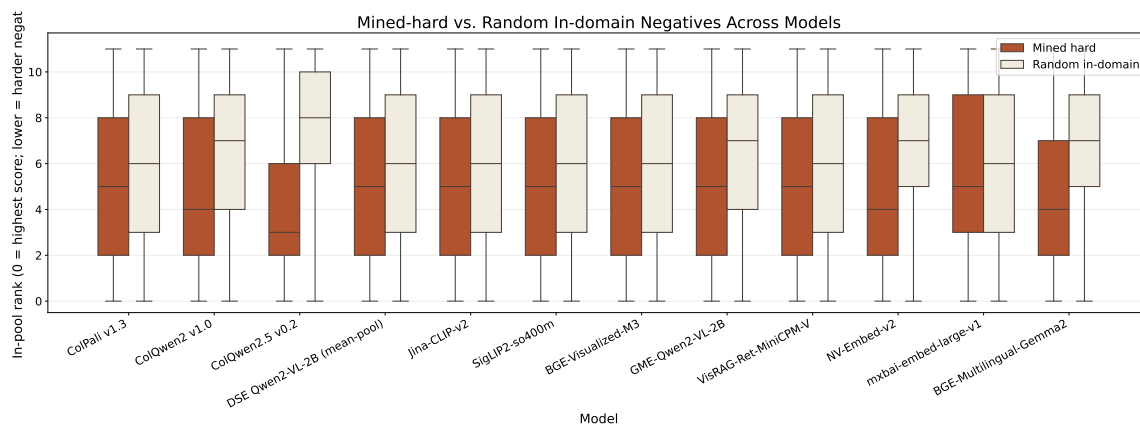


Figure 10: In-pool rank distributions for mined hard negatives and random in-domain negatives across the 12 auxiliary retrievers.

hard-negative fine-tuning experiment in the main paper.

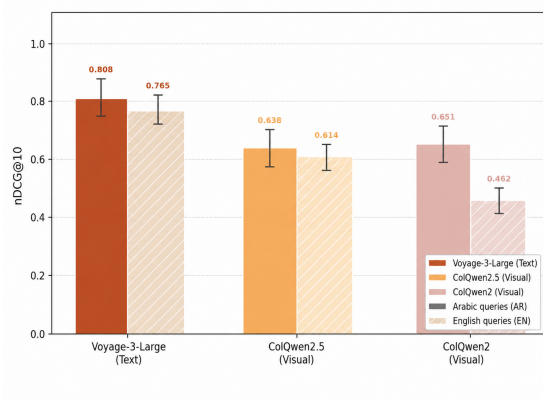


Figure 11: nDCG@10 on 200 stratified evaluation queries for native Arabic queries and NLLB-translated English queries.

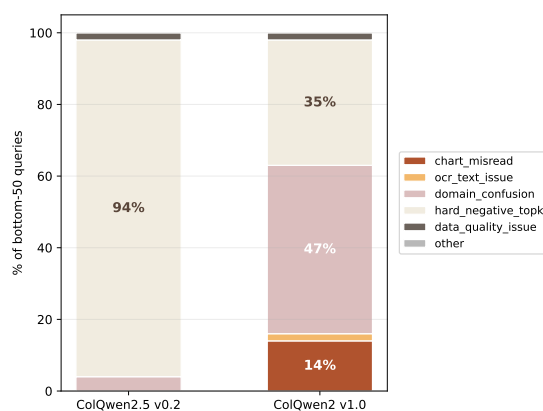


Figure 12: Failure-mode distribution for the two strongest visual retrievers.

Q#46 (bar_charts, rank=12, nDCG@10=0.0000) - ColQwen2.5 v0.2
 failure: domain_confusion
 رسم بياني الأراضي الزراعية العربية 1982 | جداول

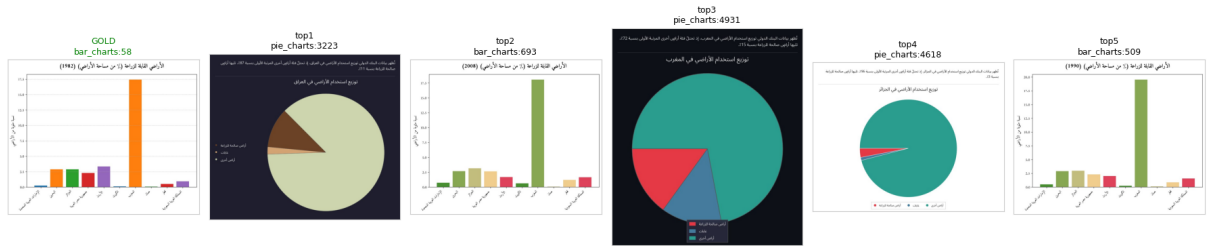


Figure 13: Failure panel 1: hard-negative confusion on bar charts. Query: *Arabic agricultural land statistics 1982*. The gold document is ranked twelfth; higher-ranked documents concern the same topic but different years or countries.

Q#8 (tables, rank=10, nDCG@10=0.0000) - ColQwen2.5 v0.2
 failure: hard_negative_topk
 كم تبلغ مساحة تايوان وعدد سكانها

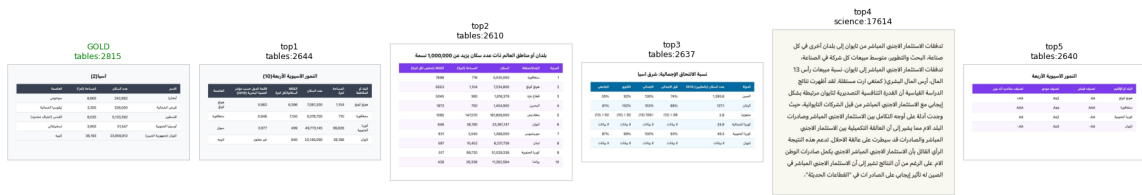


Figure 14: Failure panel 2: data-quality issue on tables. Query: *How large is Taiwan and what is its population?.* The gold table is weakly informative for the query, while higher-ranked tables are more directly relevant.

Q#44 (tables, rank=12, nDCG@10=0.0000) - ColQwen2.5 v0.2
 failure: domain_confusion
 تطور القطاع الصناعي الكويت نمطية اقتصاد

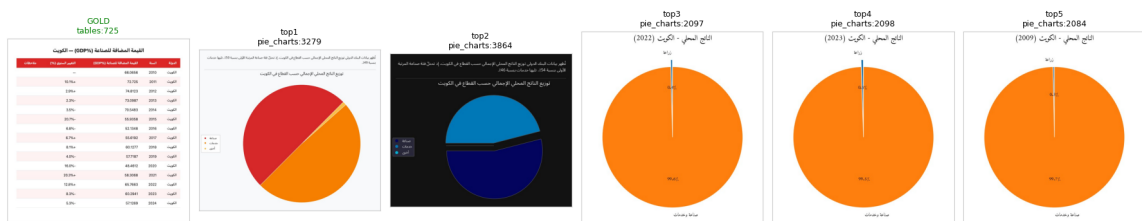


Figure 15: Failure panel 3: hard-negative confusion on tables. Query: *Evolution of Kuwait's industrial sector economic percentages*. The gold document is ranked twelfth; higher-ranked documents concern related industries but different countries or sectors.