DEAM: Diverse Embeddings for Arabic Multimodal Retrieval

Anonymous ACL submission

Figure 1: Overview of DEAM. Source documents are collected from nine Arabic domains, covering both visual-heavy categories such as tables and charts, and prose-heavy categories such as Wikipedia, news, books, science, and government documents. Pages are rendered with an HTML + Playwright pipeline to preserve Arabic shaping and bidirectional layout. Each document is paired with Arabic queries generated and refined into *direct*, *comparative*, and *long-tail* intents, then reviewed by human annotators under a two-of-three consensus protocol. The final release provides contrastive retrieval triplets with mined hard and easy negatives, together with evaluation-ready splits for Arabic visual document retrieval.

Abstract

Modern visual document retrieval (VDR) models remain strongly English-centric, leaving Arabic — the script of over 400 million speakers — 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 mining. The dataset comprises **18,275** rendered Arabic documents across nine domains paired with **82,727** contrastive retrieval triplets, each including five mined hard negatives and one easy negative for contrastive fine-tuning. Comprehensive zero-shot evaluation of **26 retrievers** reveals a sharp inversion of the English VDR ranking: strong multilingual text retrievers outperform the best visual retrievers on Arabic by 13.6 nDCG@10 points, with visual retrieval winning only on layout-rich news pages. DEAM establishes a foundational resource for Arabic visual document retrieval and will be released with code and benchmark splits to support research on under-served languages in 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) now map text, images, document pages, and even video into a shared semantic space, redefining the state of the art across retrieval, semantic textual similarity, clustering, and classification. On the visual document side, late-interaction encoders such as ColPali (Faysse et al., 2024a), the ColQwen family (Faysse et al., 2024c), DSE (Ma et al., 2024), and unified frameworks like VLM2Vec-V2 (Meng et al., 2025) show that a vision-language encoder can index a rendered page directly and rival OCR-based pipelines on standard benchmarks. The trend is clear: high-quality visual embeddings are displacing OCR-then-retrieve as the default for document search.

These advances, however, have been measured almost exclusively on English data, and the gains do not transfer uniformly across languages. Recent multilingual visual retrieval benchmarks reveal that even state-of-the-art visual embedders lose substantial accuracy on non-English scripts (Osmulski et al., 2025). Despite Arabic being the script of over 400 million users, there is no open Arabic visual document retrieval (VDR) resource that pairs rendered documents with human-validated queries; existing multilingual encoders can therefore claim only weak Arabic transfer, and the community has no dedicated corpus on which to fine-tune or fairly evaluate them.

Arabic also poses challenges that are easy to underestimate. The script is bidirectional, contextually shaped (initial / medial / final / isolated forms), and includes optional diacritics; rendering pipelines built on Pillow frequently produce detached or mis-shaped glyphs, silently corrupting any dataset built on top. Arabic content additionally exhibits a sharp register gap between Modern Standard Arabic and regional dialects which retrieval systems must handle.

We close this gap with DEAM, the first open Arabic visual document retrieval resource — distinct from the recent wave of Arabic OCR and instruction-tuning resources (see §2), which target transcription or generation rather than retrieval. Our contributions are:

- **A multi-domain Arabic VDR corpus** spanning nine categories: four visual-heavy (tables, bar / line / pie charts) and five prose-heavy (Wikipedia, news, books, science, government), rendered through an HTML + Playwright pipeline that preserves correct Arabic shaping.
- **A reproducible three-source query-generation pipeline** combining a vision-capable LLM, a text-only LLM, and a second vision LLM that refines queries into explicitly typed intents (*direct*, *comparative*, *long-tail*), yielding complementary surface forms across sources.
- **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** with mined hard negatives (via ColQwen2.5 (Faysse et al., 2024c) and TF-IDF), released alongside per-category subsets, an evaluation configuration with negatives embedded inline, and a corpus configuration.

The dataset, code, and benchmark splits will be released publicly upon acceptance; for review, we provide an anonymized mirror at `https://anonymous.4open.science/r/deam-paper-578C/README.md`.

2 Related Work

Multimodal and Text Embedding Models. Modern retrieval is dominated by large multilingual embedding encoders — BGE-M3 (Chen et al., 2024a), multilingual-E5 (Wang et al., 2022, 2024a), GTE-Qwen2 (Li et al., 2023a), Qwen3-Embedding (Zhang et al., 2025b), 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., 2024a) and ColQwen2.5 (Faysse et al., 2024c) via late-interaction MaxSim. These benchmarks are English-centric. MIRACL-VISION (Osmulski et al., 2025) is the closest multilingual analogue

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

Table 1: Where DEAM sits in the landscape. ✓ = supports the criterion; ✗ = does not; (✓) = partial, where Arabic is one of 18 languages in MIRACL-VISION rather than the dedicated focus. *Multi-source queries* denotes more than one independent LLM-based query generator combined with deduplication. *Human validated* denotes a human-labelled gold split. *Hard-neg mining* denotes triplets shipped with mined hard negatives. DEAM is the only Arabic-dedicated visual document retrieval resource with multi-source queries, human consensus validation, and shipped hard negatives.

— 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 (Table 1).

Arabic Document Resources: Recognition vs Retrieval. A parallel Arabic thread targets *transcription* rather than retrieval: KITAB-Bench (Heakl et al., 2025) benchmarks Arabic OCR across nine domains, 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., 2023), Peacock (Alwajih 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 standard for training dense retrievers: Doc2Query (Nogueira et al., 2019), InPars (Bonifacio et al., 2022), and Promptagator (Dai et al., 2023) popularised prompting LLMs as few-shot query generators with filtering. Prior work generates from *text*; DEAM extends this paradigm to the *visual* setting with a multi-source pipeline (Claude vision + Claude text + GPT-4o typed refinement) and 21-annotator human consensus.

3 DEAM Dataset

DEAM is built through a five-stage human-in-the-loop pipeline (Figure 2). Human effort is concentrated at the *endpoints*: 21 community contributors curate the source corpus (stage i), then independently review every LLM-generated query under a two-of-three consensus protocol on Argilla (Argilla, 2024) (stage iv). The middle three stages — rendering, query generation, and hard-negative mining — are fully automated.

3.1 Source Collection

Arabic source material is gathered from publicly available content across nine domains. *Prose*: Wikipedia, news, books, science, and government. *Visual*: tables, bar charts, line charts, and pie charts. After deduplication, content-quality filtering, and Unicode normalisation, the corpus comprises 18,275 cleaned records (Table 2). All sources are permissively licensed.

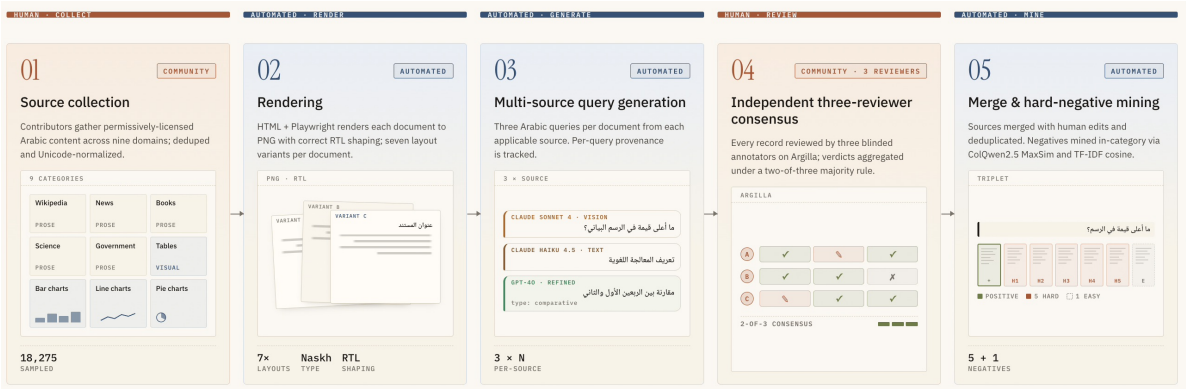

Figure 2: The five-stage construction pipeline behind DEAM: source collection, rendering, query generation, human validation, and hard-negative mining.

Category	Docs	Queries (triplets)
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.

3.2 Rendering

Documents are rendered to PNG via an HTML + Playwright[1] pipeline rather than Pillow: Pillow-based Arabic rendering produces detached or misshaped glyphs in our pilot, while the HTML renderer guarantees correct contextual shaping via the browser's text-shaping engine. We use Naskh-style fonts, set `direction: rtl`, dynamically size each page to fit its content, and apply seven layout variants per document (cross product of two fonts, two font sizes, and several backgrounds) for visual diversity. Charts and tables are rendered from structured numerical data with a randomised chart style. A 500-document spot check found no visible defects; remaining layout-quality failures are caught by the human reviewers (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, 2024) 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, 2024) consumes the underlying text content of prose documents under the same diverse-query schema — 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, producing a refined trio with explicit typing — *direct* (main subject), *comparative* (max/min/trend), and *long-tail* (specific entity or number) — adding 28,086 typed queries in total. Exact prompts are reproduced in Appendix A. After deduplication on a normalised 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 community contributors served as annotators on a self-hosted Argilla (Argilla, 2024) deployment; **every test record was independently reviewed by three of them**, blind to each other's verdicts (18,275 annotation rows). Each annotator supplied two judgements per record: an *image-quality* verdict (good / bad rendering — records flagged `bad_rendering` by majority are dropped, catching failures the automated spot check missed) and a per-query verdict of *approve*, *edit*, or *reject*, with an optional edited string. The *edit* option lets reviewers rescue near-miss queries rather than discard useful information, seeding the `human_edited` provenance flag in the merged release. We aggre-

[1] https://playwright.dev/

gate per-query verdicts under a **two-of-three majority rule**; queries with no majority are dropped. Annotator-quality statistics (Cohen's κ, Fleiss' κ, per-annotator consensus agreement) appear in §4.

3.5 Merge, Deduplication, and Hard-Negative Mining

We merge all generated and human-edited queries, deduplicate on a normalised 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., 2024c) — a late-interaction VL retriever — scores each query against in-category document images using MaxSim:

$$\mathrm{MaxSim}(q, d) \;=\; \sum_i \max_j \langle \bar{q}_i, \bar{d}_j \rangle, \qquad (1)$$

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 campaign produced 18,275 annotations from 21 reviewers (mean 870 per reviewer) aggregated under a two-of-three majority rule (Table 3). Reviewer pairs reach **75.2%** pairwise agreement; chance-corrected scores are lower (Cohen's $\kappa = 0.184$, Fleiss' $\kappa = 0.210$) — a known artefact of highly skewed label distributions (Feinstein and Cicchetti, 1990), since the consensus approval rate is $\approx 89\%$ and chance agreement itself is therefore high. Per-category κ ranges from 0.060 (news, mostly *approve*) to 0.325 (bar charts, where *edit* is more common). At consensus, **89.5%** of judged query positions are approved (3,760/273/169 approve/edit/reject across 4,209 positions; 7 records have no majority). A separate `image_quality` verdict drops **332 records** with majority `bad_rendering`, concentrated in line charts

Metric	Value
Annotators	21
Annotation rows logged	18,275
Test queries (consensus)	3,000
Pairwise verdict agreement (raw)	75.2%
Cohen's κ (pairwise, macro)	0.184
Fleiss' κ (multi-annotator)	0.210
Approve / Edit / Reject @ consensus (%)	89.5 / 6.5 / 4.0
Approve / Edit / Reject @ raw votes (%)	81.4 / 8.2 / 10.4
Records dropped (`bad_rendering`)	332

Table 3: Annotation quality summary for the human-validated dataset

Source	N	Appr.	Edit	Rej.
Claude	15,275	81.4%	8.2%	10.4%
GPT-4o-refined gold	3,000	100.0%	0.0%	0.0%

Table 4: Verdict rate by query source in the human-reviewed set.

(45.5% of reviewed records) and science documents (32.5%); Figure 3 shows the per-category breakdown.

Table 4 breaks the verdict distribution down by query source: 15,275 reviewed instances come from Claude generation, while 3,000 are GPT-4o-refined typed queries that constitute the gold test split.

4.2 Query Length and Vocabulary Diversity

After human review the final query set spans a range of formulations across sources (whitespace tokenisation after NFKC normalisation and Arabic diacritic removal). Claude-only queries are concise with a median of 8 tokens; GPT-4o-refined queries are slightly longer (median 10, $p90{=}12$), reflecting their explicit *direct* / *comparative* / *long-tail* typing; human-edited queries are the longest ($\bar{x} = 10.7$, $p90{=}14$), as annotators typically added clarifying details rather than shortening. Pairwise unigram Jaccard similarity across sources averages 0.35–0.46, indicating that no single style dominates the reviewed set; the highest overlap appears in bar charts (Claude–human Jaccard $= 0.65$), where chart titles strongly constrain the natural query wording.

5 Experimental Setup

We accompany DEAM with a zero-shot retrieval benchmark spanning **26 baselines** — 17 text-only and 9 visual document retrievers — and a publicly released harness for reproducing them. The goal of this paper is to characterize where the field stands on Arabic visual document retrieval today;

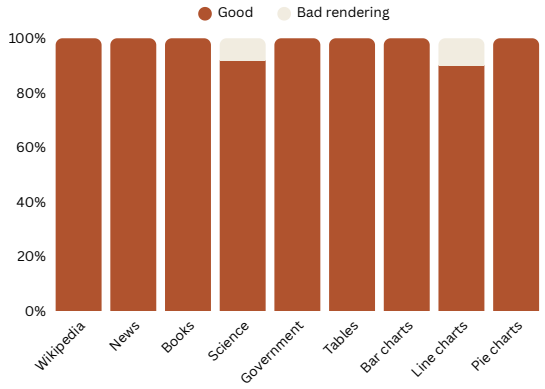

Figure 3: Distribution of the consensus `image_quality` verdict per category. Line charts and science have the highest `bad_rendering` rates.

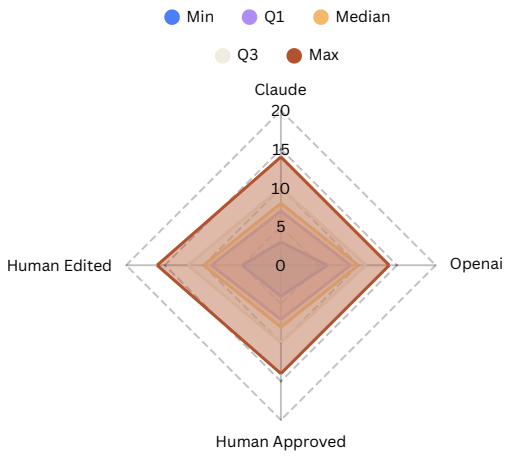

Figure 4: Query token length distribution per source group. Whiskers cover 1.5 IQR. GPT-4o-refined queries are slightly longer than Claude originals; human-edited queries are the longest.

advancing the state of the art through fine-tuning is left to follow-on work. Accordingly, **no model is fine-tuned on DEAM** in this paper. The training and validation splits, together with their mined hard negatives, will be released to enable that follow-on work.

5.1 Test Split and Evaluation Protocol

Retrieval is evaluated on the **3,000 queries** of the human-validated test split that have a valid positive image and a 2-of-3 majority consensus verdict (§3.4). For every query, each retriever ranks all **18,275 documents** in the corpus; we report nDCG@10 (primary, rank-sensitive), Recall@{1, 5, 10}, and MRR@10, each with a 95% bootstrap confidence interval over 1,000 resamples of the query set. The test split is concentrated in three high-volume categories — news (n=1,434), bar charts (n=809), and tables (n=748) — with small books (n=3) and government (n=6) tails.

Family	Evaluated baselines
Lexical	BM25 (Robertson and Zaragoza, 2009).
Open multilingual	BGE-M3 (Chen et al., 2024a); mE5-large (Wang et al., 2022); GTE-Qwen2-7B (Li et al., 2023a); Jina-v3 (Sturua et al., 2024); Nomic-v2-MoE (Nussbaum and Duderstadt, 2025); Qwen3-Emb-8B (Zhang et al., 2025b); 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., 2024a); ColQwen2-v1.0 (Faysse et al., 2024c); ColQwen2.5-v0.2 (Faysse et al., 2024c).
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); ColSmolVLM-alpha (Faysse et al., 2024a).

Table 5: Summary of the zero-shot baselines evaluated on DEAM.

5.2 Baselines

We choose baselines to span the full design space of contemporary retrievers along four axes: *modality* (text vs. visual), *language coverage* (English, multilingual, Arabic-specific), *access* (open weights vs. closed API), and *architecture* (sparse lexical, dense single-vector, late-interaction multi-vector, CLIP-style dual-encoder). Table 5 catalogs the 26 systems grouped by family.

Implementation details. Text retrievers use each model card's recommended pooling and query / document prompt prefixes; visual retrievers follow each reference implementation's encode procedure on the rendered PNG page images at native resolution. Commercial endpoints (OpenAI, Cohere, Voyage) are queried with each provider's default retrieval settings, and all corpus embeddings are cached. Per-baseline encoder revisions, prefixes, hardware, and runtimes are documented in Appendix B.

6 Results and Discussion

6.1 Main Retrieval Results

Tables 6 and 7 report zero-shot performance for all 26 baselines on the gold test split. The headline finding is a sharp inversion of the ranking observed on English ViDoRe-style benchmarks:

Model	nDCG@10	R@1	R@10	MRR
Lexical baseline				
BM25 (Robertson and Zaragoza, 2009)	0.617	0.541	0.697	0.592
Open multilingual				
BGE-M3 (Chen et al., 2024a)	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., 2023a)	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., 2025b)	0.831	0.755	0.909	0.806
Llama-Nemotron-8B (Babakhin et al., 2025)	0.223	0.156	0.299	0.199
Arabic-specific				
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 (Omartificial-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
Commercial API				
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 6: Zero-shot text retrievers on the DEAM test split.

text-only retrievers substantially outperform visual retrievers on Arabic documents, with the best text system (Voyage-3-Large, nDCG@10 = 0.833) leading the best visual system (ColQwen2.5-v0.2, nDCG@10 = 0.697) by 13.6 absolute nDCG points.

Why does text beat visual on Arabic? We attribute the inversion to two compounding factors. First, the VL backbones underlying most visual retrievers (CLIP, SigLIP, BLIP, Qwen2-VL) were pretrained on heavily English-dominant image–caption corpora; their visual text understanding for Arabic glyphs and right-to-left layout is correspondingly weak. Second, the `title + text_content` field consumed by text retrievers is clean Unicode, while a visual retriever must recover the same information *through* the rendered glyphs, paying an avoidable accuracy tax on Arabic shaping and direction. Late-interaction visual retrievers with Qwen2.5-VL backbones (ColQwen2.5) close part of this gap but do not match a strong multilingual text encoder on the global test set.

6.2 Per-Domain Performance

Figure 5 reports per-category nDCG@10 with 95% bootstrap CIs for the strongest text retriever (Voyage-3-Large) and the top visual retrievers (ColQwen2.5, ColQwen2) on the three high-volume test categories: news, bar charts, and tables. The per-domain results show that the aggregate text-over-visual trend hides important variation. ColQwen2.5 leads on news (0.959 vs. 0.889), likely benefiting from layout cues such as mastheads and bylines. Voyage-3-Large dominates on bar charts (0.715 vs. 0.275), where retrieval depends on reading axis labels and numerical content. Tables are intermediate: Voyage remains ahead (0.881 vs. 0.770), but the smaller gap suggests that table structure provides useful visual cues for late-interaction models.

Model	nDCG@10	R@1	R@10	MRR
Late-interaction multi-vector				
ColPali-v1.3 (Faysse et al., 2024a)	0.227	0.148	0.323	0.198
ColQwen2-v1.0 (Faysse et al., 2024c)	0.652	0.554	0.758	0.619
ColQwen2.5-v0.2 (Faysse et al., 2024c)	**0.697**	**0.612**	**0.787**	**0.669**
Single-vector VLM				
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
CLIP-style multilingual VLM				
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
Other visual retrievers				
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., 2024a)	0.029	0.020	0.040	0.025

Table 7: Zero-shot visual document retrieval results on the DEAM test split. **Bold** marks the best visual retriever.

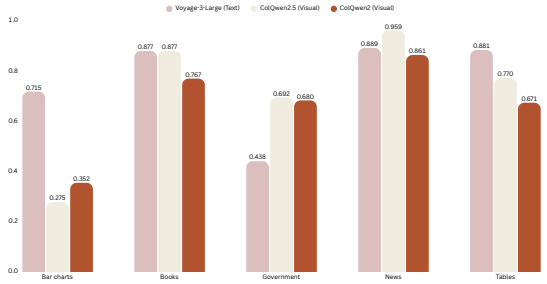

Figure 5: Per-domain nDCG@10 for the strongest text retriever and the top two visual retrievers. Error bars show 95% bootstrap confidence intervals.

6.3 Cross-Lingual Generalization

A natural shortcut for serving Arabic users with English-trained models is to translate the query into English and retrieve against the Arabic corpus. We test this on 200 stratified-sampled test queries translated with NLLB-200-distilled-600M, run against the unchanged Arabic corpus by the top text and top two visual retrievers as shown in (Table 8).

Native Arabic queries consistently outperform

Model	AR queries	EN (NLLB)	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 8: Cross-lingual nDCG@10 on 200 stratified test queries. Native Arabic queries consistently outperform NLLB-translated English queries, especially for visual retrievers.

translated ones, with the gap ranging from 4.2 nDCG points for Voyage to 18.9 points for ColQwen2. This validates the dedicated Arabic-language design of DEAM: translation-as-bridge is not a viable shortcut, particularly for visual retrievers whose VL backbones already see limited Arabic supervision.

6.4 Discussion

Text-retriever ranking: open multilinguals close the commercial gap; Arabic-specific embedders lag. Voyage-3-Large, Qwen3-Embedding-8B, and Cohere Embed-v4.0 cluster within 0.005 nDCG at the top, with the strongest *open* model (Qwen3-Embedding-8B) matching the commercial APIs; second-tier open multilinguals (mE5, GTE-Qwen2, Jina-v3, Nomic-MoE) trail by 2–3 points, while OpenAI `text-embedding-3-large` underperforms this tier by ∼6 points, suggesting its multilingual mix is under-weighted for Arabic. By contrast, Arabic-specific text embedders (AraGemma, SILMA, Arabic-Triplet-Matryoshka, GATE-AraBERT-v1) sit in a 0.65–0.72 band — reflecting STS-style training data rather than retrieval supervision, not a verdict on Arabic modelling — making them prime candidates for fine-tuning on DEAM's triplets.

Mining quality and residual failures. We cross-validate the released hard negatives by ranking, for each test query, the gold positive against 5 mined and 6 random in-domain negatives under 12 independent retrievers: mined negatives rank significantly lower (harder) than random for **all 12 models**, $p<0.001$ in 10 of 12 (Wilcoxon signed-rank, one-sided). A manual taxonomy of the bottom-50 ColQwen2.5 queries further shows that 94% of residual failures are hard-negative top-K confusion within the correct domain — precisely the regime contrastive fine-tuning on DEAM's mined triplets is designed to address. Full breakdowns appear in Appendices D and F.

Path forward for Arabic visual retrieval. The combination of (i) text-beats-visual on the headline number, (ii) visual-beats-text on layout-rich news, and (iii) hard-negative confusion as the dominant residual failure points to a clear direction: Arabic VDR will benefit most from VL backbones with native Arabic glyph and RTL understanding plus contrastive fine-tuning on the mined hard-negative triplets we release. We leave this to follow-on work, with the training and validation splits of DEAM made available to enable it.

7 Conclusion

We presented DEAM, the first open Arabic visual document retrieval resource: 18,275 rendered Arabic documents across nine domains paired with 82,727 contrastive retrieval triplets and a 3,000-query gold test split. Benchmarking 26 zero-shot retrievers reveals a sharp inversion of the English visual-document retrieval ranking — strong multilingual text retrievers beat the best visual retrievers on Arabic by 13.6 nDCG@10 points, with visual retrieval winning only on layout-rich news — and the dominant residual failure of the best visual retriever is hard-negative confusion within the correct domain, exactly the regime that contrastive fine-tuning on the released triplets is designed to address. By releasing the corpus, the training and validation splits with mined hard negatives, and the full benchmark harness, we enable a new line of Arabic visual document retrieval fine-tuning and evaluation research.

8 Limitations

A central limitation of this paper is that we do not fine-tune any retriever on DEAM; the reported numbers characterise off-the-shelf models only. The strongest open Arabic-specific embedders we evaluate (AraGemma, SILMA, GATE-AraBERT-v1, Arabic-Triplet-Matryoshka-V2) sit in the 0.65–0.72 nDCG@10 band — substantially below strong multilingual encoders — and no Arabic-dedicated *visual* document retriever yet exists at all. Closing this gap by contrastive fine-tuning a strong (V)LM backbone on the released training triplets is exactly the follow-on work that DEAM is designed to enable, and we leave it as a clear direction for future research.

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*.

Hassan Al-Homoud, Ahmed A. Ibrahim, Mohammed Al-Jubran, Faisal Al-Otaibi, Yousef Alharbi, Dauren 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.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohammed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report.

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. In *Proceedings of SIGIR*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. 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.

W. Bruce Croft, Donald Metzler, Trevor Strohman, and 1 others. 2010. *Search Engines: Information Retrieval in Practice*, volume 520. Addison-Wesley Reading.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *Proceedings of ICLR*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024a. ColPali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024b. ColPali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024c. 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 Sohail, M. Ranjan, R. Hossam, G. Ahmed, M. El-Geish, O. Maher, Z. Shen, F. Khan, and S. H. Khan. 2025. KITAB-Bench: A comprehensive multi-domain benchmark for Arabic ocr and document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Khalil Hennara, Muhammad Hreden, Mohamed M. Hamed, Ahmed Bastati, Zeyad 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, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Juncai He, and 1 others. 2023. AceGPT, localizing large language models in Arabic. *arXiv preprint arXiv:2309.12053*.

Vladimir Karpukhin, Barlas Oğuz, 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 EMNLP*.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR*.

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, Sibo 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. 2023a. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. 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. *arXiv preprint arXiv:2406.11251*.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2025. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*.

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 Sibaee, 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*.

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. In *arXiv:1904.08375*.

Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *arXiv preprint arXiv:2502.07972*.

Omartificial-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 and 1 others. 2025. MIRACL-VISION: A multilingual multimodal document retrieval benchmark. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

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, Andrius Skaburskas, David Michael Jones, Charles Pierse, Rafael Esposito, Joseph Trengrove, Etienne Dilocker, and Bob van Luijt. 2026. IRPAPERS: 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*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

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. 2024a. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. 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 Wenhu Chen. 2024. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision (ECCV)*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025a. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yanzhao Zhang, Dingkun Long, Mingxin Li, Pengjun Xie, and 1 others. 2025b. 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.

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 6: Compact version of the 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 7: Compact version of the Claude Haiku 4.5 text prompt used for prose categories: Wikipedia, news, books, science, and government. The model receives document text rather than the rendered image because prose images do not add information beyond their textual content.

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). All prompts are presented in English for consistency and readability. Figures 6 and 7 show the query-generation prompts used for visual and prose documents, respectively.

The GPT-4o refinement prompt was issued in English and is reproduced verbatim in Figure 8. The `{n}` placeholder in the Claude prompts is filled with `3` at call time (three queries per document); the `{text}` placeholder in the Haiku prompt is filled with the first 800 characters of the document's `text_content`. The original Arabic prompts are available verbatim in the released code repository.

Model	nDCG@10	R@1	R@5	R@10	MRR
BM25 (Croft et al., 2010)	0.617	0.541	0.656	0.697	0.592
BGE-M3 (Chen et al., 2024b)	0.780	0.697	0.827	0.865	0.753
mE5-large (Wang et al., 2024b)	0.804	0.732	0.842	0.878	0.780
GTE-Qwen2-7B (Li et al., 2023b)	0.798	0.717	0.844	0.880	0.771
Jina-Embeddings-v3 (Sturua et al., 2024)	0.806	0.728	0.852	0.883	0.781
Nomic-Embed-v2-MoE (Nussbaum and Duderstadt, 2025)	0.792	0.708	0.839	0.879	0.764
Qwen3-Embedding-8B (Zhang et al., 2025a)	0.831	0.755	0.875	0.909	0.806
Llama-Embed-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-Matryoshka-V2 (Nacar and Koubaa, 2025)	0.651	0.545	0.703	0.766	0.614
SA-Retrieval-Embed-0.2B (Nacar, 2025)	0.390	0.314	0.428	0.474	0.363
AraGemma-Embed-300m (Omartificial-Intelligence-Space, 2025)	0.718	0.603	0.781	0.840	0.679
SILMA-Embed-Matryoshka (Soliman et al., 2024)	0.660	0.549	0.719	0.775	0.624
OpenAI `embed-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 9: Full metrics for all text retrievers on the DEAM test split. 95% bootstrap CIs (1,000 resamples) are available in the released results CSV.

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/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 8: Compact version of the GPT-4o typed refinement prompt used for the four visual categories. The model receives the rendered image, document type, and three candidate queries, then returns three Arabic queries with explicit *direct*, *comparative*, and *long_tail* labels under a strict JSON schema.

B Baseline Protocols

This appendix records the exact protocol used for each zero-shot baseline so the numbers in Tables 6 and 7 are fully reproducible.

Text retrievers. All open dense text retrievers use the model card's recommended pooling (mean pooling unless the card specifies CLS) and prompt prefixes. Where applicable: BGE-M3 uses no prefix; mE5-large prepends `"query: "` and `"passage: "`; GTE-Qwen2-7B-instruct uses its instruction template with the retrieval task description; Qwen3-Embedding-8B uses its task-instruction prefix; Jina-v3 uses the `retrieval.query` / `retrieval.passage` task labels; Nomic-Embed-v2-MoE uses `search_query:` / `search_document:`. BM25 is implemented via Pyserini with the default Arabic analyser; Arabic-specific embedders (AraGemma, SILMA, Arabic-Triplet-Matryoshka, GATE-AraBERT-v1, SA-Retrieval) follow their published preprocessing pipelines. Closed-source endpoints (OpenAI `text-embedding-3-large`, Cohere Embed-v4.0, Voyage-3-Large) are called with each provider's default retrieval settings; corpus embeddings are cached. All dense retrievers compute cosine similarity over L2-normalized vectors, rank exhaustively against the 18,275-document corpus, and report metrics with 1,000 bootstrap resamples.

Visual retrievers. ColPali-v1.3, ColQwen2-v1.0, and ColQwen2.5-v0.2 are run via the `colpali-engine` library at their published encoder revisions, with MaxSim scoring over L2-normalized patch / token vectors. GME-Qwen2-VL-2B and DSE Qwen2-VL-2B are mean-pooled single-vector encoders; the latter follows the DSE reference implementation. Jina-CLIP-v2 and SigLIP2 use the standard CLIP protocol (separate image / text towers, cosine similarity). BGE-Visualized-M3, VisRAG, and ColSmolVLM-alpha follow their respective reference codebases. Image inputs are the rendered PNG page images at native resolution.

Hardware. Visual encoders were run on a single RTX 5090 (32 GB) at fp16 / bf16 in roughly 30–60 minutes per encoder for the 18,275-document cor-

Model	nDCG@10	R@1	R@5	R@10	MRR
ColPali-v1.3 (Faysse et al., 2025)	0.227	0.148	0.265	0.323	0.198
ColQwen2-v1.0 (Faysse et al., 2024b)	0.652	0.554	0.704	0.758	0.619
ColQwen2.5-v0.2 (Faysse et al., 2024b)	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-so400m (Tschannen et al., 2025)	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-Ret-MiniCPM (Yu et al., 2025)	0.017	0.011	0.020	0.024	0.014
ColSmolVLM-alpha (Faysse et al., 2024b)	0.029	0.020	0.031	0.040	0.025

Table 10: Full metrics for all visual retrievers on the DEAM test split.

pus. Text encoders were run on the same hardware (or via API where applicable); BM25 indexing finishes in minutes on CPU.

C Detailed Per-Baseline Results

Tables 9 and 10 report all five metrics (nDCG@10, R@1, R@5, R@10, MRR@10) for the text and visual baselines respectively. The main-paper tables omit R@5 for brevity.

D Mining Quality: Cross-Model Validation

This appendix expands the mining-quality validation summarized in §6.4. For each of the 2,660 test queries, every evaluated model ranks a candidate pool consisting of the gold positive, the 5 mined hard negatives (from `neg_ids`, originally mined by ColQwen2.5), and 6 random in-domain negatives (fixed seed 42). Per-query mean ranks of mined-hard vs random candidates are compared with a one-sided Wilcoxon signed-rank test (H_1: mined $<$ random).

Table 11 reports the mean in-pool rank for both candidate types, the rank difference (random $-$ mined; positive $=$ mined is harder), and the Wilcoxon p-value. Across all **12 evaluated retrievers** — including five visual models, the mining oracle (ColQwen2.5), and four independent multilingual / NV text embedders that were never used for mining — mined hard negatives rank lower than random. Ten of twelve are significant at $p<0.001$.

E Cross-Lingual Generalization: Visual View

Figure 12 visualises the cross-lingual finding reported as Table 8 in the main text. The visual view makes two observations easier to see than the table: (i) native Arabic queries dominate machine-translated English queries on all three evaluated retrievers, and (ii) the magnitude of the gap is strongly model-dependent — almost negligible for ColQwen2.5 (-2.4 nDCG) but catastrophic for ColQwen2 (-18.9 nDCG).

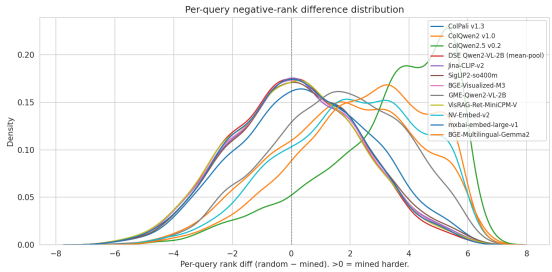

Figure 9: Per-query rank difference (random $-$ mined) across the 12 evaluated retrievers. Positive values indicate mined negatives are harder. The mining oracle (ColQwen2.5) shows the largest shift; the effect persists in independent retrievers.

F Qualitative Failure-Mode Analysis

This appendix expands the qualitative error analysis summarized in 6.4. For each of the two strongest visual retrievers (ColQwen2.5 and ColQwen2) we manually inspected the 50 bottom-ranked test queries by per-query nDCG@10 and tagged each failure into a six-class taxonomy.

Failure-mode definitions. *Hard-negative top-K*: a same-domain, same-topic distractor outranked the gold positive in the top-K; this is the failure mode contrastive fine-tuning on DEAM's mined negatives directly targets. *Domain confusion*: the retrieved document is from the wrong domain entirely. *Chart misread*: the model matched a chart's visual layout but missed the metric or label encoded in the chart. *OCR/text issue*: the rendered text was poorly readable. *Data quality issue*: the gold document was empty, mislabelled, or otherwise weakly annotated. *Other*: residual cases that do not fit the above categories.

Model	Mined rank	Random rank	Δ (random $-$ mined)	Wilcoxon p
ColQwen2.5-v0.2 *(oracle)*	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 (mean)	5.47	5.55	0.08	0.033

Table 11: Cross-model mining validation. Mean in-pool rank of mined hard negatives versus random in-domain negatives across the candidate pool (1 positive + 5 mined + 6 random), aggregated over 2,660 test queries. Positive Δ indicates mined negatives rank lower (are harder); the Wilcoxon test is one-sided (H_1: mined $<$ random). All 12 models show the expected direction; 10 of 12 are significant at $p<0.001$.

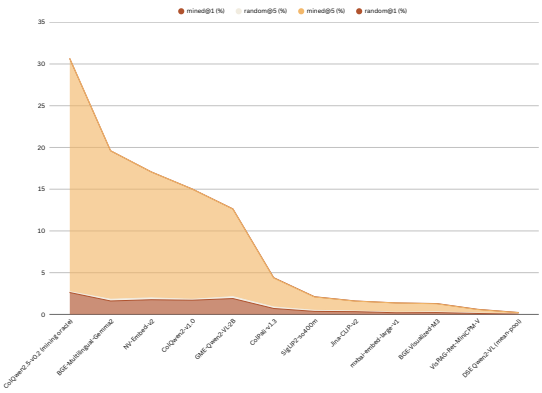

Figure 10: Fraction of mined hard negatives that land in each retriever's full-corpus top-1 / top-5 / top-10 (red), versus the same fraction for random in-domain negatives (blue). For ColQwen2.5, 43% of mined negatives appear in the top-10 against the full 18,275-document corpus, versus 0.3% for random in-domain negatives.

Example failure panels. Figures 14, 15, and 16 show three illustrative failure panels under ColQwen2.5, covering one hard-negative top-K failure on bar charts, one data-quality failure on tables, and one hard-negative top-K failure on tables. Each panel shows the Arabic query (top), the gold positive document, and the five top-ranked retrieved documents. The dominant pattern is that the retrieved documents are visually and topically near-identical to the gold positive but disagree on a single numeric or named-entity detail — exactly the bottleneck that contrastive fine-tuning on DEAM's mined hard-negative triplets is designed to address.

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/text issue	0	1 (2%)
Data quality issue	1 (2%)	0
Other	0	1 (2%)
Total	50	51

Table 12: Failure-mode counts on the bottom-50 nDCG@10 queries of ColQwen2.5 and ColQwen2. The strongest visual retriever is dominated by within-domain hard-negative confusion; the older model also suffers cross-domain confusion and chart misreads.

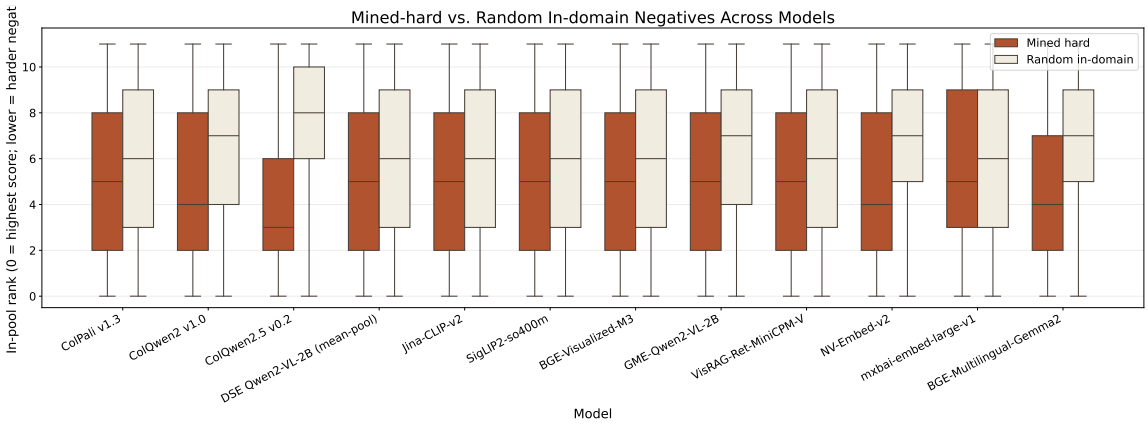

Figure 11: In-pool rank distributions for mined hard negatives and random in-domain negatives across the 12 evaluated retrievers. Brown denotes mined hard negatives and light beige denotes random in-domain negatives. Across models, the mined distributions are consistently shifted toward lower ranks, indicating that the mined negatives are substantially more challenging than randomly sampled in-domain negatives. The size of this shift generally tracks each model's overall retrieval strength.

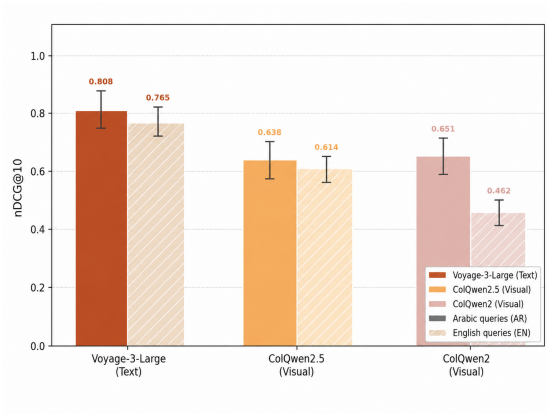

Figure 12: nDCG@10 on 200 stratified test queries when the query is issued in native Arabic (left bar per model) versus when it is machine-translated to English via NLLB-200-distilled-600M (right bar per model). The Arabic corpus is unchanged. Error bars are 95% bootstrap CIs.

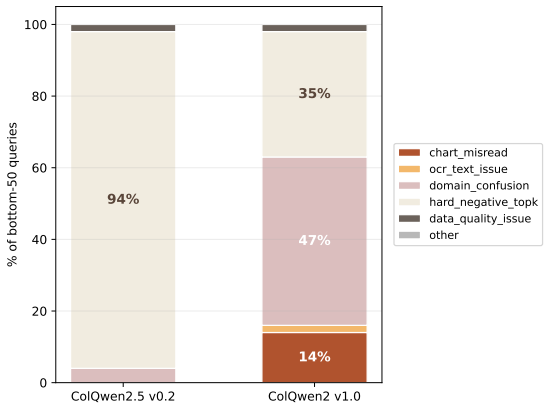

Figure 13: Failure-mode distribution (percentages) on the bottom-50 nDCG@10 queries for the two strongest visual retrievers. ColQwen2.5 is dominated by hard-negative confusion within the correct domain; ColQwen2 also suffers domain confusion.

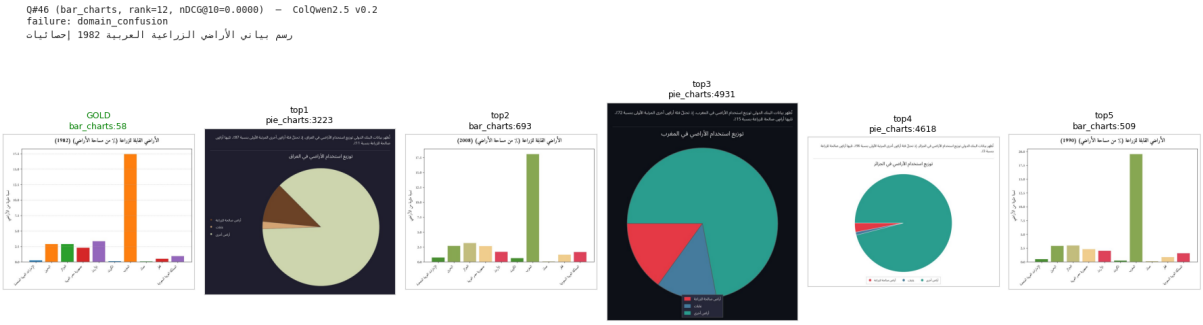

Figure 14: Failure panel #1 — hard-negative top-K on bar charts. Query (Arabic): *Arabic agricultural land statistics 1982*. Gold positive at rank 12; the four higher-ranked retrieved documents are bar charts on the same topic with different years or countries.

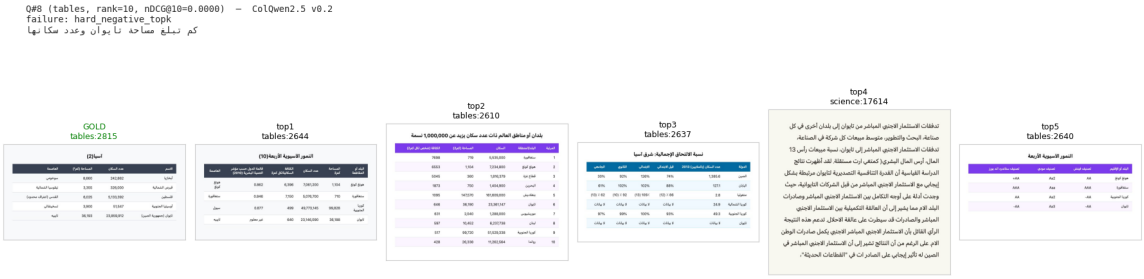

Figure 15: Failure panel #2 — data-quality issue on tables. Query (Arabic): *How large is Taiwan and what is its population?*. Gold positive at rank 10; inspection reveals the gold table is weakly populated for this query, while the higher-ranked tables are more directly informative.

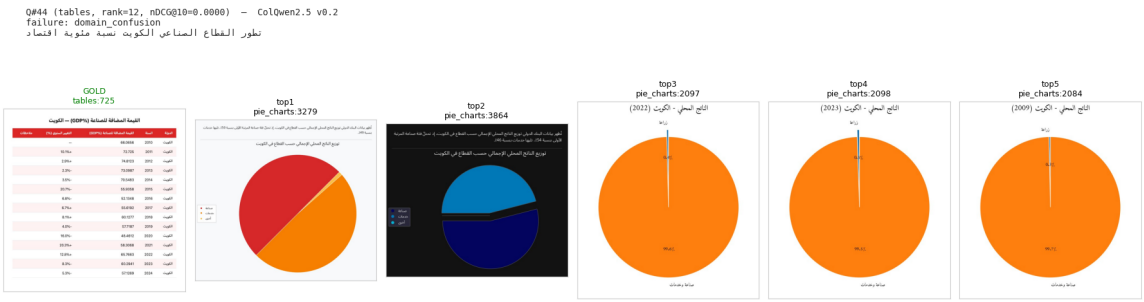

Figure 16: Failure panel #3 — hard-negative top-K on tables. Query (Arabic): *Evolution of Kuwait's industrial sector economic percentages*. Gold positive at rank 12; the higher-ranked documents are tables on the same topic with different countries or different industrial sectors.