RAVEL: SEPARATING ROMANIZATION COLLISIONS IN MULTILINGUAL DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Romanization erases script cues that separate names, allowing a faithful variant of one entity to share its surface form with another identity. We formulate script robustness as two coupled obligations: reconnect relevant variants and keep identity-colliding passages outside the retrieved neighborhood. The Fracture–Collapse Protocol (FCP) evaluates benchmark-wide noisy Recall@100 together with candidate-micro collision intrusion over a model-independent population, and connects retrieval behavior to neighborhood geometry, source-disjoint noise, real input-method errors, generated answers, and serving cost. RAVEL realizes this constraint with BGE-M3 semantics, a 12M-parameter byte/character form channel, and identity-labeled collision negatives, while retaining one vector per query and passage. Across MKQA, mMARCO, and XOR-Retrieve, it gains 6.9–8.3 Recall@100 points over BGE-M3 and 3.5–3.7 points over the same-backbone invariance control, with 0.3–0.4 points of clean regression and 32 ms p95 latency. On MKQA, RAVEL reaches 69.7 noisy recall at 1.8% candidate intrusion, compared with 66.2 and 4.1% for invariance; its 6.9-point source-disjoint advantage persists, and fixed-generator wrong-entity answers fall from 4.8% to 3.3%. These results establish identity separation as a co-equal criterion for training and selecting script-robust single-vector retrievers.

1 INTRODUCTION

Romanization is lossy with respect to identity. Converting a name to Latin script can preserve how it is typed while removing the characters that distinguish one person, place, or term from another. The same ambiguity enters through partial or wrong-mode input-method conversion before retrieval and through OCR residue or mixed-script passages after indexing. Because a colliding passage can remain topically plausible, the error can silently propagate from dense retrieval into a generated answer (Karpukhin et al., 2020; Lewis et al., 2020; Guu et al., 2020; Shi et al., 2024).

The usual robustness objective solves only one side of this problem: it pulls a clean item and its Romanized view together. In our same-backbone control, that objective raises MKQA noisy Recall@100 from 62.8 to 66.2 while increasing candidate-micro collision intrusion from 2.4% to 4.1%. Relevant variants reconnect, but ambiguity-matched false friends also move closer. We call the lost access to relevant passages *neighborhood fracture* and the lost boundary between colliding identities *semantic collapse*.

This paper makes the two effects a coupled retrieval target. FCP measures benchmark-wide noisy recall and top-100 entry from a fixed, model-independent population of distinct-identity Romanization collisions. Neighborhood margins reveal the representation change behind both outcomes; source-disjoint OCR and code-switch families, real input-method errors, per-model adaptive edits, and fixed-generator answers test whether the same ordering survives changes in surface process and evaluation level. All dense systems emit one vector per query and passage and share the FAISS-HNSW configuration, while translation and late interaction define adjacent systems envelopes.

RAVEL (Romanization-Aware Vector Embeddings for Lexical Ambiguity) encodes the two relations explicitly. BGE-M3 supplies semantic relevance, a four-layer byte/character channel supplies form evidence, and identity-labeled collision negatives repel strings that become confusable after Romanization. The channels combine into one normalized 1024-dimensional vector, preserving a single-vector

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

passage index and ANN search. Empirical-bank augmentation and cross-view invariance provide same-backbone controls for noise exposure and positive-only alignment.

The evidence links objective, geometry, retrieval, and answer identity. Across MKQA, mMARCO, and XOR-Retrieve, RAVEL improves noisy Recall@100 over BGE-M3 by 6.9–8.3 points and over invariance by 3.5–3.7 points, with 0.3–0.4 clean-point regression. On ambiguous MKQA queries, it expands the positive–false-friend margin from 0.06 under invariance to 0.17 and lowers rank flips from 28.9% to 17.2%. The gain persists under excluded source families and real IME input; with a fixed generator, MKQA noisy exact match rises by 5.4 points over BGE-M3 while wrong-entity answers fall from 4.8% to 3.3%.

The contributions are:

1. FCP, a collision-conditioned protocol that places benchmark-wide noisy relevance beside fixed-population candidate intrusion and traces both to neighborhood geometry.

2. RAVEL, a single-vector retriever that assigns form recovery and identity separation to distinct training signals without changing the ANN index structure.

3. A controlled three-benchmark evidence chain—same-backbone controls, excluded source families, real IME input, generated answers, ablations, and measured serving cost—that establishes the retrieval and downstream consequence of preserving Romanization identity boundaries.

2 RELATED WORK

Romanized and mixed-script retrieval. Transliterated search studies spelling expansion and lexical matching (Prabhakar et al., 2021). Lost in Transliteration adapts neural retrievers with transliterated relevance training (Chari et al., 2025), and TransliCo aligns representations across scripts (Liu et al., 2024). Both motivate the positive relation in RAVEL: native-script and Romanized views of the same target should agree. FCP adds the complementary evaluation object—a fixed population that shares the normalized Romanized form but not the identity—and RAVEL makes that relation repulsive. The BGE-M3+invariance control instantiates positive-only cross-view alignment on the same backbone, data, and ANN stack, so the central comparison changes the collision-aware design rather than the encoder family or serving path.

Surface-noise and code-switch robustness. Character and byte encoders reduce tokenization sensitivity (Xue et al., 2022; Clark et al., 2022; Tay et al., 2022), while contrastive adaptation improves dense retrieval under misspellings and OCR corruption (Belinkov & Bisk, 2018; Campos et al., 2023; Michail et al., 2025). Code-mixed training exposes tension between language alignment and relevance matching (Do et al., 2024), and CSR-L and CS-MTEB broaden code-switch retrieval evaluation (Zeng et al., 2026). These approaches model difficult surfaces; the relation supervised here is identity-specific. Conventional relevance negatives are semantically competing passages, whereas a RAVEL collision negative is selected because its normalized Romanized token sequence matches and its benchmark identity differs. The evaluator, miner, and geometric diagnostic therefore operate on the same relation.

Multilingual retrieval and serving regimes. Single-vector models include mDPR, mContriever, mE5, mAggretriever, BGE-M3, and E5-Mistral (Zhang et al., 2021; Izacard et al., 2022; Wang et al., 2024b; Lin et al., 2023; Chen et al., 2024; Wang et al., 2024a); multilingual suites include MKQA, mMARCO, MIRACL, and MTEB (Longpre et al., 2021; Bonifacio et al., 2021; Zhang et al., 2023; Muennighoff et al., 2023). FCP augments their usual relevance reporting with collision-conditioned identity entry. BGE-M3 variants carry the causal comparison because they share backbone, training envelope, and ANN configuration. Translation and ColBERT-X (Nair et al., 2022) provide systems context under their complete online paths, with quality, latency, throughput, and index memory reported together.

3 COLLISION-CONSTRAINED SCRIPT ROBUSTNESS

Let $f(q)$ and $f(d)$ denote normalized single-vector query and passage embeddings. For retrieval metric M, neighborhood fracture is the script-noise robustness gap

$$\Delta_{\text{rob}} = M(f, \mathcal{Q}_{\text{clean}}, \mathcal{D}) - M(f, \mathcal{Q}_{\text{noisy}}, \mathcal{D}). \tag{1}$$

The surface process may Romanize a query, leave OCR residue in a passage, mix scripts within a span, stop IME conversion early, or apply at most three bank-supported local edits. Relevance labels, corpus membership, model weights, and ANN state remain fixed.

Semantic collapse uses a model-independent candidate population. For a text item x with a designated mention, let $\rho(x)$ be that mention's lexicon-supported Romanized token sequence after Unicode normalization and benchmark alias closure, and let $y(x)$ be its benchmark entity or normalized target identity. A same-identity item with matching ρ is a genuine view; a matching-ρ, different-y item is a collision. For each ambiguity query $q \in \mathcal{A}$, $\mathcal{E}(q)$ contains the corpus passages with such a distinct-identity collision. Candidate construction finishes before retrieval. Duplicate passages collapse by corpus identifier, each retained colliding mention maps to one identity, and $\mathcal{A}$ contains exactly the queries with a labeled target and at least one distinct-identity candidate. Aliases of one entity are therefore positives, while unresolved multi-entity passages never enter $\mathcal{E}(q)$.

Let $r_f(q, d)$ be the rank of d in the full corpus under model f, with score ties broken by corpus identifier. A candidate outside the top k contributes zero. The ambiguity intrusion rate is

$$\text{FP}_{\text{amb}}(f) = \frac{\sum_{q \in \mathcal{A}} \sum_{d \in \mathcal{E}(q)} \mathbb{1}[r_f(q, d) \le k] \mathbb{1}[y(d) \ne y(q)]}{\sum_{q \in \mathcal{A}} |\mathcal{E}(q)|}. \tag{2}$$

Every eligible candidate occupies one slot for every model; retrieval changes only the indicator $r_f(q, d) \le k$, and a candidate outside the top k contributes zero. The denominators are therefore fixed at 24,600, 31,400, and 18,700 slots for MKQA, mMARCO, and XOR, respectively. FCP reports candidate-micro FP_{amb} beside language-macro Recall@100. These are complementary benchmark-wide and ambiguity-slice estimands unless a table explicitly identifies a shared paired slice. Clean quality, neighborhood margin, and serving cost complete the profile.

4 RAVEL: ALIGN GENUINE VARIANTS, REPEL ROMANIZATION COLLISIONS

RAVEL assigns distinct capacity to the two sides of FCP: BGE-M3 supplies semantic relevance, while a four-layer byte/character Transformer with 12M parameters recovers form variants. Queries and passages share the complete dual-channel encoder. Mean-attention pooling produces the form representation, which is projected to the semantic channel's 1024-dimensional space. The channels are independently ℓ_2-normalized, combined, and normalized again:

$$z(x) = \text{norm}(\alpha \ \text{norm}(f_{\text{sem}}(x)) + (1 - \alpha) \ \text{norm}(W f_{\text{form}}(x))) \,. \tag{3}$$

Development selection gives $\alpha = 0.7$. Inference computes both channels once and stores a single 1024-dimensional passage vector, leaving HNSW structure and top-k unchanged. The semantic backbone has 568M parameters and the complete encoder has 580M; the 12M increment is isolated to the form channel rather than the index.

For in-batch passages $\mathcal{B}$ and temperature τ, relevance learning uses

$$\mathcal{L}_{\text{rel}} = -\sum_i \log \frac{\exp(z(q_i)^\top z(d_i^+)/\tau)}{\sum_{d \in \mathcal{B}} \exp(z(q_i)^\top z(d)/\tau)}. \tag{4}$$

Cross-view robustness is $\mathcal{L}_{\text{rob}} = \sum_i \|z(q_i) - z(\tilde{q}_i^+)\|_2^2$. For a genuine view $\tilde{q}_i^+$ and a collision-matched text item $\tilde{x}_i^-$ with $\rho(\tilde{x}_i^-) = \rho(q_i)$ and $y(\tilde{x}_i^-) \ne y(q_i)$, separation is applied to the final retrieval vector z:

$$\mathcal{L}_{\text{amb}} = \sum_i \max(0, \|z(q_i) - z(\tilde{q}_i^+)\|_2 - \gamma^+) + \sum_i \max(0, \gamma^- - \|z(q_i) - z(\tilde{x}_i^-)\|_2). \tag{5}$$

The complete objective is $\mathcal{L} = \mathcal{L}_{\text{rel}} + 0.5\mathcal{L}_{\text{rob}} + 0.3\mathcal{L}_{\text{amb}}$, with $\gamma^+ = 0.1$ and $\gamma^- = 0.4$. Each relevance anchor is paired with 2 cross-view positives and 4 ambiguity negatives. The miner refreshes

once per epoch over training queries and passages, applies the same ρ/y test as FCP, and excludes every evaluation query, passage, and label. Because queries and passages share the encoder, $z(\tilde{x}_i^-)$ uses the same path regardless of which text side supplies the collision.

The corruption bank derives confusion statistics or transformation rules from OCR, transliteration, code-switch, and IME source families. Source-disjoint evaluation excludes every upstream corpus or lexicon family used to fit those rules, rather than holding out only instances. Local adversarial search chooses the bank-supported sequence of at most three character operations that most reduces the relevant-passage margin. Training uses AdamW, learning rate 5×10^{-5}, batch size 256, 10 epochs, and eight A100 80GB GPUs with CUDA 12.4 and PyTorch 2.3.1. During epochs 1–5, optimization updates the form channel and projection while the semantic branch remains frozen; epochs 6–10 update both branches.

Two same-backbone controls isolate the design. *BGE-M3+bank* applies the identical training corpus, bank sampling, optimizer steps, and seeds without cross-view alignment, form channel, or ambiguity negatives. *BGE-M3+invariance* adds the identical cross-view objective but retains one BGE-M3 encoder and omits the form channel and ambiguity negatives. Their contrast with RAVEL measures the value of explicit form capacity and semantic separation.

5 EXPERIMENTS

The experiments answer three linked questions: Does positive-only cross-view alignment admit more colliding identities? Does explicit separation improve identity and relevance together? Does the resulting ordering persist across source families, answer generation, and a measured single-vector serving envelope?

5.1 FCP CONSTRUCTION AND ESTIMATION

All non-adaptive evaluation manifests are fixed before model scoring. Romanization views use normalized lexicon pairs and spelling variants; OCR views use source-derived confusions and independently rescanned passages; code-switch views use observed span-length and language-pair statistics; and real-IME views preserve partial conversion, wrong-mode, or adjacent-key input. The 50% combined condition independently perturbs query and passage surfaces. Adaptive evaluation is the explicit exception: for each model, it selects at most three bank-supported edits that minimize that model's relevant-minus-false-friend margin. Upstream corpus or lexicon families, rather than individual strings, define source-disjointness. Labels and corpus membership remain unchanged; Appendix A records the surface, construction, and comparison unit for every condition.

MKQA-Retrieval and mMARCO use Arabic, Hindi, Bengali, Russian, Chinese, Japanese, Korean, and Thai (Longpre et al., 2021; Bonifacio et al., 2021); XOR-Retrieve is aggregated separately over Arabic, Bengali, Japanese, Korean, Russian, and Telugu (Asai et al., 2021). The benchmark-wide tests contain 8,000, 55,840, and 4,257 queries. Held-out OCR, code-switch, four-language combined, and adaptive slices contain 4,800, 6,400, 4,000, and 8,000 queries. The MKQA ambiguity slice contains 2,240 queries and 24,600 fixed candidate slots; Appendix B gives the complete inventory.

The source-disjoint test has 3,200 Romanization-heavy queries and 64,000 OCR/code-switch passages; the real-IME slice has 4,000 Hindi and CJK queries. Answer evaluation fixes Llama-3.1-8B-Instruct, greedy decoding, and one retriever-independent prompt, passage order, context budget, truncation rule, and generation limit across the 1,600 MKQA and 1,200 XOR queries. Retrieved passages are presented in rank order. EM, F1, and identity matching use the same normalized benchmark answer aliases; a wrong-entity answer contains a collision-identity alias but no gold-identity alias, and WEA is query-micro over each fixed subset.

Baselines are Unicode normalization+mContriever, ByT5-Retriever, CANINE-Retriever, BGE-M3, E5-Mistral-7B, the two same-backbone controls, and online translation to English followed by mContriever. The BGE-M3 variants and component removals support attribution; cross-architecture rows measure the broader quality–cost context. ColBERT-X appears as a late-interaction systems comparison. Every single-vector model uses FAISS-HNSW with $M = 32$, efSearch$= 128$, and top 100. Serving measurements include query encoding and ANN lookup on one A100 80GB and an

Intel Xeon Gold 6430 host, batch size one, 16 clients, 1,000 warm-up queries, and 10,000 timed queries. Translation and ColBERT-X retain their complete online paths.

Learned variants report five-seed means; frozen off-the-shelf systems are evaluated once on the same manifests. Recall uncertainty uses 10,000 language-stratified paired-query resamples and recomputes the language-macro statistic. $\mathrm{FP}_{\mathrm{amb}}$ is the candidate-micro estimator in Equation (2), averaged across training seeds for learned variants; Appendix B converts each displayed mean rate to a nearest-integer equivalent event count on the fixed denominator rather than pooling the same candidate slots five times. Paired intervals support the recall contrasts; collision-intrusion and wrong-entity-answer tables report descriptive rates on their declared fixed populations.

5.2 Alignment Recovers Surface Form but Can Lose Identity

Table 1: The central fracture–collapse comparison. Recall@100 is benchmark-wide; FP is fixed-population ambiguity intrusion (%) on the corresponding benchmark slice.

	MKQA		mMARCO		XOR	
Method	R@100	FP	R@100	FP	R@100	FP
BGE-M3	62.8	2.4	62.1	2.7	59.4	3.1
E5-Mistral	62.4	2.2	61.8	2.5	58.7	3.0
BGE-M3+bank	65.4	3.1	65.8	3.2	62.4	3.6
BGE-M3+invariance	66.2	4.1	66.7	4.2	63.5	4.6
RAVEL	69.7	1.8	70.4	1.9	67.2	2.2

Cross-view invariance exposes the trade-off on all three benchmarks: it raises noisy recall over BGE-M3 while increasing intrusion to 4.1/4.2/4.6%. RAVEL instead improves recall beyond invariance and reduces intrusion by 2.3/2.3/2.4 points. On MKQA's fixed 24,600 candidate slots, the displayed rates correspond to 590 BGE-M3, 1,009 invariance, and 443 RAVEL rate-equivalent intruding slots. These candidate-micro values are independent of whether a relevant passage also remains in the top 100.

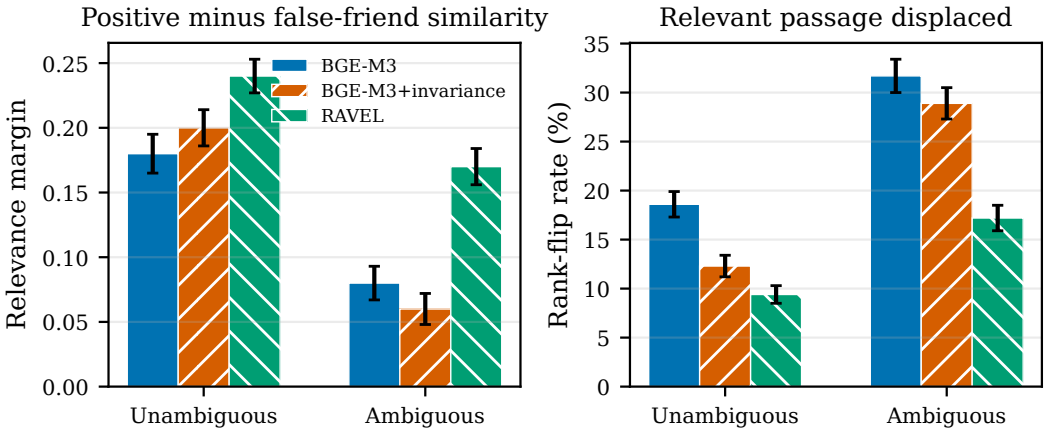

Figure 1: Geometry on 2,240 ambiguity-matched and 2,240 unambiguous queries. Paired-bootstrap 95% intervals are shown. Invariance pulls both positives and false friends closer; RAVEL restores the positive–false-friend margin and reduces rank flips. Exact values appear in Appendix D.

The geometry explains the divergent rates. On ambiguous inputs, BGE-M3 has a 0.08 positive–false-friend margin; invariance increases positive similarity but compresses the margin to 0.06. RAVEL expands it to 0.17 and lowers rank flips from 28.9% to 17.2%. Both robust variants improve the unambiguous slice, whereas only collision separation protects the ambiguous neighborhood.

Table 2: Neighborhood-fracture evidence: benchmark-wide Recall@100 and clean-to-noisy gap. MKQA and mMARCO are eight-language macro means; XOR is an independent six-language macro mean. All single-vector rows share the ANN configuration.

	MKQA			mMARCO			XOR			ms	
Method	Cln	Noi	Gap	Cln	Noi	Gap	Cln	Noi	Gap	p50	p95
Unicode+mContr.	81.2	54.4	26.8	78.6	51.8	26.8	74.3	48.1	26.2	14	22
ByT5	79.8	56.7	23.1	77.4	55.2	22.2	72.1	51.4	20.7	28	42
CANINE	80.4	58.2	22.2	78.1	57.4	20.7	73.6	54.2	19.4	24	38
BGE-M3	83.4	62.8	20.6	81.7	62.1	19.6	77.8	59.4	18.4	18	28
E5-Mistral	84.7	62.4	22.3	83.2	61.8	21.4	79.4	58.7	20.7	82	124
BGE-M3+bank	83.3	65.4	17.9	81.5	65.8	15.7	77.6	62.4	15.2	18	29
BGE-M3+invariance	83.2	66.2	17.0	81.4	66.7	14.7	77.4	63.5	13.9	19	29
Translate (ctx.)	82.1	63.0	19.1	80.4	63.8	16.6	76.2	60.8	15.4	142	218
RAVEL	83.1	69.7	13.4	81.3	70.4	10.9	77.4	67.2	10.2	21	32

5.3 COLLISION SEPARATION ALSO RESTORES RELEVANT NEIGHBORHOODS

RAVEL gains 6.9, 8.3, and 7.8 noisy-recall points over BGE-M3, with paired intervals [6.1, 7.8], [7.4, 9.1], and [6.8, 8.9]. More importantly for attribution, it exceeds the same-backbone invariance control by 3.5/3.7/3.7 points, with intervals [2.7, 4.3], [2.9, 4.5], and [2.8, 4.7]. E5-Mistral's clean scores are highest, but its 20.7–22.3-point script-noise gaps show that clean capacity alone does not preserve the relevant neighborhood. RAVEL's gain costs 0.3–0.4 clean points and 3 ms p95 relative to BGE-M3.

Table 3: Recall@100 in hard conditions. OCR and Switch are held-out real subsets; Comb. is the fixed four-language 50% query–document condition; Adv. applies a separately optimized attack of at most three bank-supported edits to each model.

Method	OCR	Switch	Comb.	Adv.	p50 ms	p95 ms
Unicode+mContr.	47.2	55.8	39.4	42.1	14	22
ByT5	52.8	58.1	44.7	48.4	28	42
CANINE	54.1	59.7	46.2	49.8	24	38
BGE-M3	58.4	65.2	49.8	53.6	18	28
E5-Mistral	57.1	64.8	48.3	51.4	82	124
BGE-M3+bank	61.3	67.1	53.6	56.7	18	29
BGE-M3+invariance	62.7	68.3	55.4	58.1	19	29
Translate (ctx.)	59.6	66.4	51.2	50.7	142	218
RAVEL	65.8	71.4	58.6	61.2	21	32

RAVEL improves over BGE-M3 by 7.4 points on real OCR, 6.2 on real code-switching, 8.8 under combined noise, and 7.6 under the per-model adaptive attack. The complete design adds 3.1–3.2 points over invariance in every column; the component ablation below assigns the separate form and collision signatures. Combined query–passage corruption remains the lowest absolute-recall condition and the largest RAVEL–BGE-M3 gap.

5.4 Excluded Sources and Generated Answers Preserve the Mechanism

Table 4: Source-disjoint and real-IME evaluation. Xfer combines excluded OCR/code-switch source families; IME covers Hindi and CJK input settings. R@100 and FP are paired on each fixed slice.

	Source-disjoint		Real IME		
Method	R@100	FP	R@100	FP	p95 ms
BGE-M3	54.8	3.6	64.6	3.2	28
E5-Mistral	53.9	3.4	64.0	3.0	124
BGE-M3+bank	58.1	3.8	67.1	3.4	29
BGE-M3+invariance	60.2	5.0	68.9	4.8	29
Translate (ctx.)	60.8	3.2	68.0	2.9	218
RAVEL	61.7	2.4	71.0	2.1	32

RAVEL's source-disjoint Recall@100 exceeds BGE-M3 by 6.9 points and invariance by 1.5, with paired intervals [5.8, 8.0] and [0.6, 2.4]. On real IME input it adds 6.4 points over BGE-M3 and reaches 71.0 recall at 2.1% FP; invariance reaches 68.9 at 4.8% FP. Appendix C adds nDCG@10, clean regression, Hindi/CJK splits, and fixed-denominator ledgers.

Table 5: Retrieval-to-answer consequence with a fixed generator. EM and F1 are answer scores; Cont. is gold-answer containment in retrieved context; WEA is wrong-entity answer rate.

Data	Retriever	Clean EM	Noisy EM	Noisy F1	Cont.	WEA
MKQA	BGE-M3	45.8	34.2	43.7	55.9	4.8
	BGE-M3+bank	45.7	36.6	46.2	59.1	5.2
	BGE-M3+invariance	45.6	37.8	47.3	60.5	7.1
	RAVEL	45.5	39.6	49.4	63.1	3.3
XOR	BGE-M3	40.9	30.6	39.2	52.1	5.5
	BGE-M3+bank	40.8	32.9	41.5	55.4	5.9
	BGE-M3+invariance	40.7	34.1	42.8	57.2	8.0
	RAVEL	40.6	35.9	44.8	59.8	3.8

The retrieval ordering carries into generated answers. On MKQA, RAVEL adds 5.4 noisy EM points over BGE-M3 and reduces wrong-entity answers by 1.5 points. Invariance raises noisy EM but also raises WEA to 7.1%; RAVEL reaches the highest EM at 3.3% WEA. XOR shows the same distinction, with a 5.3-point EM gain over BGE-M3 and 1.7-point WEA reduction.

5.5 Both Sides of the Constraint Are Necessary

Table 6: MKQA ablation. Values are eight-language means under the same five-seed protocol.

Variant	Clean	Noisy	Gap	FP (%)
RAVEL	83.1	69.7	13.4	1.8
− form channel	83.4	63.1	20.3	2.3
− ambiguity negatives	82.4	67.8	14.6	4.7
− empirical bank	83.0	64.2	18.8	2.1
− adversarial search	83.2	68.1	15.1	1.9
− cross-view alignment	83.3	65.4	17.9	2.0

The one-factor removals match the objective. Without the form channel, noisy recall falls by 6.6 points while intrusion changes by 0.5 points. Without ambiguity negatives, most recall remains, but intrusion rises from 1.8% to 4.7%. Bank exposure and cross-view alignment account for additional recall, and adaptive-search training contributes 1.6 points. Sensitivity over α, form capacity, ambiguity weight, and edit budget retains the same recall–intrusion region around the selected setting (Appendix E).

5.6 Single-Vector Serving Cost

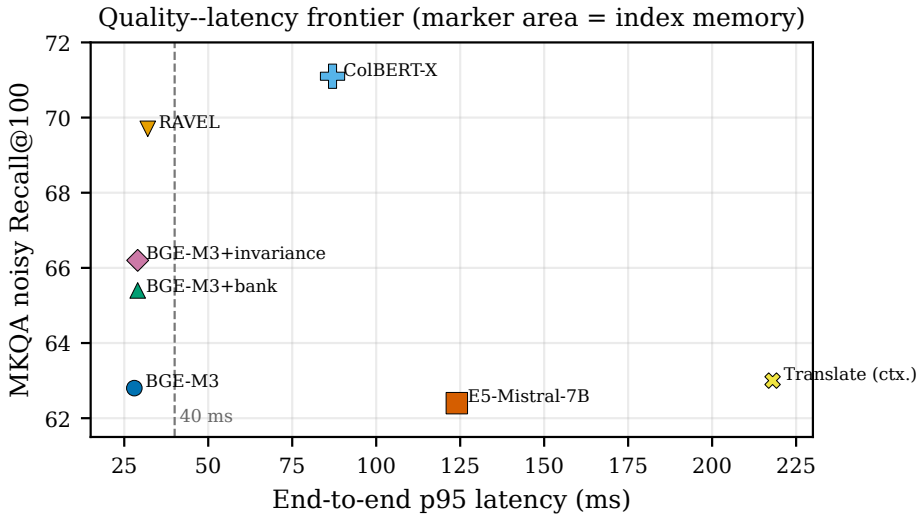

Figure 2: MKQA noisy Recall@100 versus end-to-end p95 latency; marker area denotes index memory. ColBERT-X provides the adjacent late-interaction envelope. Among the primary configurations in the figure, RAVEL is the highest-quality single-vector point below 40 ms. Exact latency, throughput, and memory values appear in Appendix F.

ColBERT-X reaches 71.1 Recall@100 at 87 ms p95 with 246.0 GiB of index memory. The selected RAVEL configuration delivers 69.7 at 32 ms with 42.6 GiB and 164 queries/s. Among the primary configurations in Figure 2, it is the highest-recall single-vector point below 40 ms; the translate path is lower-recall and adds online translation latency.

6 Discussion

Positive-pair invariance is an incomplete model-selection target for script robustness. It recovers relevant passages, but its geometry also raises the similarity of Romanization-colliding identities; on MKQA, choosing it by noisy recall alone would accept 4.1% candidate intrusion. FCP exposes that movement over a fixed population, and RAVEL supplies the repulsive relation that changes the outcome to 69.7 recall and 1.8% intrusion. The ablation signatures—form capacity primarily affects recall, while collision negatives primarily affect intrusion—connect the behavior to the two objective terms.

Serving measurements separate the representation contribution from a richer retrieval stack. RAVEL keeps one 1024-dimensional vector and the BGE-M3 HNSW structure, reaching 69.7 MKQA noisy recall at 32 ms p95. Translation and late interaction occupy 218 and 87 ms p95 systems envelopes, respectively. Source-disjoint, IME, and fixed-generator results preserve the same ordering across surface sources and from retrieval into answer identity.

The evidence defines a concrete operating regime: source-derived noise and lexicon-supported collisions in the evaluated Arabic, Indic, Cyrillic, and CJK settings; benchmark targets and normalized aliases as identity labels; top-100 single-vector retrieval; and one fixed generator interface. Within that regime, fracture–collapse separation repeats across three benchmarks, excluded source families, real IME input, neighborhood geometry, and answer identity. Combined query–passage corruption remains the hardest measured condition and still yields the largest RAVEL–BGE-M3 gain.

7 Conclusion

Romanization robustness requires two geometric relations: genuine variants must reconnect, and distinct identities with the same Latin surface must remain apart. FCP measures both with noisy

recall and fixed-population candidate intrusion; RAVEL implements them through form-sensitive alignment and identity-labeled separation in one vector. Its 6.9–8.3 Recall@100 gains over BGE-M3, 1.8% MKQA intrusion, 32 ms p95 latency, and lower wrong-entity answer rates show that relevance recovery and identity preservation can improve together. Script-robust retrievers should therefore be trained and selected on both obligations, rather than recall alone.

ETHICS STATEMENT

Ambiguous Romanization can route multilingual users to passages about the wrong person, place, or event while producing superficially plausible downstream answers. Reporting fixed-population candidate intrusion and wrong-entity answer rates makes this risk inspectable across languages. All evaluated resources use benchmark labels and derived corruption statistics under their source terms; the analysis does not infer sensitive attributes or identify individual users.

AI USE STATEMENT

Generative language-model assistance was used for document organization, technical exposition, consistency checks across definitions, tables, and references, presentation-level naming, and this submission-preparation audit. No generative language model executed experiments or generated underlying research data or numerical results.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the estimands, model, objectives, controls, evaluation construction, answer protocol, and serving setup. The appendix records evaluation counts, fixed FP denominators, recall intervals, transfer splits, sensitivity, systems costs, and the source-family separation used by the reported tests.

REFERENCES

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 547–564, 2021. doi: 10.18653/v1/2021.naacl-main.46. URL https://aclanthology.org/2021.naacl-main.46/.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A multilingual version of the MS MARCO passage ranking dataset, 2021. URL https://arxiv.org/abs/2108.13897.

Daniel Campos, ChengXiang Zhai, and Alessandro Magnani. Noise-robust dense retrieval via contrastive alignment post training. *arXiv preprint arXiv:2304.03401*, 2023. URL https://arxiv.org/abs/2304.03401.

Andreas Chari, Iadh Ounis, and Sean MacAvaney. Lost in transliteration: Bridging the script gap in neural IR. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2900–2905, 2025. doi: 10.1145/3726302.3730226. URL https://doi.org/10.1145/3726302.3730226.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022. doi: 10.1162/tacl_a_00448. URL https://aclanthology.org/2022.tacl-1.5/.

Junggeun Do, Jaeseong Lee, and Seung-won Hwang. ContrastiveMix: Overcoming code-mixing dilemma in cross-lingual transfer for information retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 197–204, 2024. doi: 10.18653/v1/2024.naacl-short.17. URL https://aclanthology.org/2024.naacl-short.17/.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. REALM: Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Sheng-Chieh Lin, Amin Ahmad, and Jimmy Lin. maggretriever: A simple yet effective approach to zero-shot multilingual dense retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11688–11696, 2023. doi: 10.18653/v1/2023.emnlp-main.715. URL https://aclanthology.org/2023.emnlp-main.715/.

Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schütze. TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2476–2499, 2024. doi: 10.18653/v1/2024.acl-long.136. URL https://aclanthology.org/2024.acl-long.136/.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021. doi: 10.1162/tacl_a_00433. URL https://aclanthology.org/2021.tacl-1.82/.

Andrianos Michail, Juri Opitz, Yining Wang, Robin Meister, Rico Sennrich, and Simon Clematide. Cheap character noise for OCR-robust multilingual embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11705–11716, 2025. doi: 10.18653/v1/2025.findings-acl.609. URL https://aclanthology.org/2025.findings-acl.609/.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. Transfer learning approaches for building cross-language dense retrieval models. In *Advances in Information Retrieval*, volume 13185 of *Lecture Notes in Computer Science*, pp. 382–396. Springer, 2022. doi: 10.1007/978-3-030-99736-6_26. URL https://doi.org/10.1007/978-3-030-99736-6_26.

Dinesh Kumar Prabhakar, Sukomal Pal, and Chiranjeev Kumar. Query expansion for transliterated text retrieval. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–34, 2021. doi: 10.1145/3447649.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JtBRnrlOEFN.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, 2024a. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024b. doi: 10.48550/arXiv.2402.05672. URL https://arxiv.org/abs/2402.05672.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou', Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Qingcheng Zeng, Yuheng Lu, Zeqi Zhou, Heli Qi, Puxuan Yu, Fuheng Zhao, Hitomi Yanaka, Weihao Xuan, and Naoto Yokoya. Code-switching information retrieval: Benchmarks, analysis, and the limits of current retrievers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13055–13071, 2026. doi: 10.18653/v1/2026.findings-acl.636. URL https://aclanthology.org/2026.findings-acl.636/.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 127–137, 2021. doi: 10.18653/v1/2021.mrl-1.12. URL https://aclanthology.org/2021.mrl-1.12/.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://aclanthology.org/2023.tacl-1.63/.

A FCP Manifest and Counting Rules

FCP separates fixed-view evaluation from per-model stress testing. For every fixed condition, the manifest assigns the transformed surface, source-family identifier, severity, benchmark identity, corpus identifier, and relevance labels before any model is scored. Model evaluation changes ranks only. The adaptive condition instead generates one attacked view per model under the same operation set and edit budget, so its values compare robustness to equally constrained, model-specific search rather than performance on identical strings.

Table 7: Condition manifest. "Fixed" means the evaluated text is model-independent; source-disjointness is enforced by excluding upstream corpus or lexicon families used to fit training-bank rules.

Condition	Changed surface	Construction	Comparison unit
Romanization	Query mention	Lexicon-supported ρ sequence plus recorded spelling variants; target identity and corpus stay fixed	Fixed query
OCR	Passage span	Source-derived character confusions or independent rescan output; relevance labels stay fixed	Fixed passage
Code-switch	Assigned span	Observed span-length and language-pair statistics determine the mixed-script span	Fixed span
Combined	Query and passage	Each side is independently perturbed at the declared 50% condition	Fixed pair
Real IME	Query span	Observed partial-conversion, wrong-mode, or adjacent-key input is retained	Fixed query
Adaptive	Query span	At most three bank-supported edits minimize each model's relevant-minus-false-friend margin	Per-model query

For collision scoring, candidate construction applies ρ and y before retrieval, removes duplicate corpus identifiers and unresolved multi-identity passages, and retains a query only when a distinct-identity candidate exists. Each remaining (q, d) pair is one denominator slot. A rank above k contributes zero, which keeps the slot population identical across models. Benchmark-wide Recall@100 first averages within language and then across languages; $\mathrm{FP}_{\mathrm{amb}}$ averages candidate indicators over the fixed slot population. Source-disjoint and real-IME tables compute both quantities on the same fixed slice, whereas the main cross-benchmark table places benchmark-wide recall beside its corresponding ambiguity slice.

B Evaluation Scale and Uncertainty

Table 8: Evaluation inventory. Relevant positives count judged query–passage pairs; document entries give the indexed or perturbed pool used by the slice.

Evaluation	Languages	Queries	Relevant positives	Documents/candidates
MKQA benchmark-wide	8	8,000	8,000	21,015,324
mMARCO benchmark-wide	8	55,840	61,424	8,841,823
XOR benchmark-wide	6	4,257	5,108	25,302,117
Held-out OCR	8	4,800	5,280	96,000
Held-out code-switch	8	6,400	6,912	8,841,823
Combined 50%	4	4,000	4,400	80,000
Adaptive edits	8	8,000	8,000	21,015,324
MKQA ambiguity	8	2,240	2,240	24,600
Source-disjoint	8	3,200	3,520	64,000
Real IME	4	4,000	4,000	72,000

Table 9: Per-language query and ambiguity-candidate counts. Dashes denote languages outside XOR-Retrieve.

Language	MKQA q.	mMARCO q.	XOR q.	MKQA ambiguity candidates
Arabic	1,000	6,980	760	7,500
Hindi	1,000	6,980	–	3,200
Bengali	1,000	6,980	619	2,800
Russian	1,000	6,980	782	3,100
Chinese	1,000	6,980	–	2,400
Japanese	1,000	6,980	684	2,100
Korean	1,000	6,980	651	1,900
Thai	1,000	6,980	–	1,600
Telugu	–	–	761	–

The inventory separates benchmark scale from the collision estimator's support. In particular, the MKQA ambiguity denominator is a candidate count rather than a query count, and language-specific candidate totals sum to the same fixed 24,600 slots used by Equation (2).

Table 10: Fixed-population ambiguity ledger. Equivalent errors are the displayed candidate-micro FP rate multiplied by the fixed denominator and rounded to the nearest candidate; learned variants summarize five seed rates.

Data	Method	Equivalent errors	Denominator	FP (%)
MKQA	BGE-M3	590	24,600	2.4
	E5-Mistral	541	24,600	2.2
	BGE-M3+bank	763	24,600	3.1
	BGE-M3+invariance	1,009	24,600	4.1
	RAVEL	443	24,600	1.8
mMARCO	BGE-M3	848	31,400	2.7
	E5-Mistral	785	31,400	2.5
	BGE-M3+bank	1,005	31,400	3.2
	BGE-M3+invariance	1,319	31,400	4.2
	RAVEL	597	31,400	1.9
XOR	BGE-M3	580	18,700	3.1
	E5-Mistral	561	18,700	3.0
	BGE-M3+bank	673	18,700	3.6
	BGE-M3+invariance	860	18,700	4.6
	RAVEL	411	18,700	2.2

The fixed denominator makes the intrusion rates comparable across methods: an eligible candidate that does not enter the top 100 contributes zero for every model. Table 11 reports paired inference for Recall@100; the FP ledger reports descriptive candidate-micro rates and their event scale without treating repeated training seeds as independent candidates.

Table 11: Paired language-stratified bootstrap intervals for RAVEL's Recall@100 difference.

Setting	Comparator	Point difference	95% interval	Seed SD
MKQA noisy	BGE-M3	6.9	[6.1, 7.8]	0.54
MKQA noisy	BGE-M3+invariance	3.5	[2.7, 4.3]	0.47
mMARCO noisy	BGE-M3	8.3	[7.4, 9.1]	0.58
mMARCO noisy	BGE-M3+invariance	3.7	[2.9, 4.5]	0.51
XOR noisy	BGE-M3	7.8	[6.8, 8.9]	0.66
XOR noisy	BGE-M3+invariance	3.7	[2.8, 4.7]	0.59
Source-disjoint	BGE-M3	6.9	[5.8, 8.0]	0.71
Source-disjoint	BGE-M3+invariance	1.5	[0.6, 2.4]	0.63
Real IME	BGE-M3	6.4	[5.2, 7.5]	0.69
Real IME	BGE-M3+invariance	2.1	[1.1, 3.1]	0.61

All ten paired recall intervals remain above zero, including the comparisons with the same-backbone invariance control. Seed standard deviations are smaller than the corresponding point differences in every row.

C Source-Disjoint and IME Detail

Table 12: Full source-disjoint transfer results over five seeds. Clean regression is the method's matched clean-score change on the same query identities.

Method	R@100	nDCG@10	FP (%)	Clean regression	p95
BGE-M3	54.8	38.7	3.6	0.0	28
E5-Mistral	53.9	38.2	3.4	0.0	124
BGE-M3+bank	58.1	41.5	3.8	0.2	29
BGE-M3+invariance	60.2	43.0	5.0	0.2	29
Translate (ctx.)	60.8	43.7	3.2	1.1	218
RAVEL	61.7	44.9	2.4	0.4	32

Table 13: Real-IME Recall@100 by input setting. CJK pools Chinese, Japanese, and Korean with equal language weight.

Method	Hindi	CJK	Aggregate	FP (%)
BGE-M3	62.3	66.8	64.6	3.2
E5-Mistral	61.8	66.1	64.0	3.0
BGE-M3+bank	65.1	69.0	67.1	3.4
BGE-M3+invariance	67.0	70.8	68.9	4.8
Translate (ctx.)	66.2	69.7	68.0	2.9
RAVEL	69.4	72.5	71.0	2.1

The source-disjoint FP denominator contains 29,800 ambiguity-matched candidates; rate-equivalent errors are 1,073, 1,013, 1,132, 1,490, 954, and 715. The real-IME denominator contains 22,500 candidates, with corresponding values 720, 675, 765, 1,080, 653, and 473. These fixed-population ledgers produce the rates in Tables 12 and 13.

D Neighborhood Geometry

Table 14: Exact values behind Figure 1. Pos. and FF are noisy-query cosine similarities to the relevant passage and nearest ambiguity-matched false friend.

Slice	Method	Pos. sim.	FF sim.	Margin	Margin CI ±	Rank flip	Flip CI ±
Unambiguous	BGE-M3	0.69	0.51	0.18	0.015	18.6	1.3
	BGE-M3+inv.	0.78	0.58	0.20	0.014	12.3	1.1
	RAVEL	0.79	0.55	0.24	0.013	9.4	0.9
Ambiguous	BGE-M3	0.61	0.53	0.08	0.013	31.7	1.7
	BGE-M3+inv.	0.76	0.70	0.06	0.012	28.9	1.6
	RAVEL	0.77	0.60	0.17	0.014	17.2	1.3

E SENSITIVITY AND EFFICIENCY

Table 15: Sensitivity around the selected RAVEL configuration. Index memory remains fixed because every row stores one 1024-dimensional vector per passage.

Factor	Level	Clean	Noisy	FP	Params (M)	p50 ms	p95 ms	GiB	q/s
α	0.5	82.7	69.1	1.7	580	22	34	42.6	156
	0.6	83.0	69.5	1.7	580	21	33	42.6	160
	0.7	83.1	69.7	1.8	580	21	32	42.6	164
	0.8	83.3	68.8	2.0	580	20	31	42.6	167
Form capacity	6M	83.2	68.9	2.0	574	20	31	42.6	168
	12M	83.1	69.7	1.8	580	21	32	42.6	164
	24M	83.0	70.0	1.7	592	24	36	42.6	149
λ_{amb}	0.1	83.2	69.2	2.5	580	21	32	42.6	164
	0.3	83.1	69.7	1.8	580	21	32	42.6	164
	0.5	82.8	69.4	1.6	580	21	32	42.6	163
Edit budget	1	83.2	68.9	1.9	580	21	32	42.6	164
	3	83.1	69.7	1.8	580	21	32	42.6	164
	5	82.9	69.8	1.8	580	21	32	42.6	163

The selected 12M form channel is the knee of the tested capacity curve: the 24M variant adds 0.3 noisy-recall points while increasing p95 latency by 4 ms and reducing throughput by 15 queries/s, whereas 6M gives up 0.8 recall points. The selected setting retains the shared 42.6 GiB index allocation.

F SYSTEMS FRONTIER

Table 16: End-to-end MKQA systems frontier under identical hardware, corpus, top-100 evaluation, warm-up, and concurrency. The translation row reports the serving harness's preallocated ANN memory, not the intrinsic size of its 768-dimensional mContriever vectors.

Method	Noisy R@100	p50 ms	p95 ms	Throughput (q/s)	ANN allocation (GiB)
BGE-M3	62.8	18	28	181	42.6
E5-Mistral	62.4	82	124	44	170.4
BGE-M3+bank	65.4	18	29	177	42.6
BGE-M3+invariance	66.2	19	29	174	42.6
RAVEL	69.7	21	32	164	42.6
ColBERT-X	71.1	49	87	61	246.0
Translate (ctx.)	63.0	142	218	22	42.6

The single-vector rows share search parameters but retain their native encoders. RAVEL's 12M-parameter form channel changes query encoding time without increasing vector width or ANN allocation; ColBERT-X's multi-vector index accounts for its larger memory footprint.

G Per-Language and Severity Diagnostics

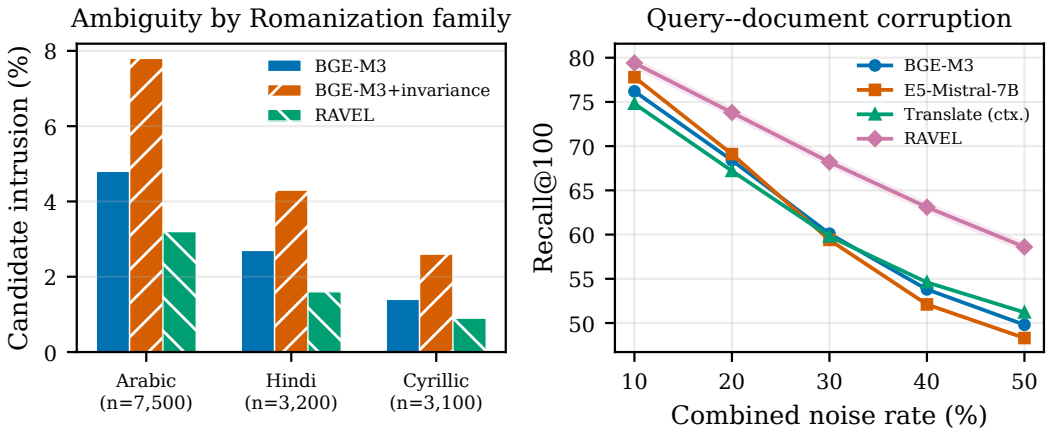

Figure 3: Left: ambiguity FP for the selected Arabic, Hindi, and Cyrillic Romanization families, with fixed candidate counts. Right: Recall@100 across combined-noise severity with RAVEL's paired-bootstrap 95% band. The 50% endpoint equals Table 3.

The family diagnostic compares identity intrusion in three distinct Romanization systems, while the severity curve shows a positive RAVEL–BGE-M3 gap at every corruption level and an 8.8-point gap at 50%. The tables below separate that selected-family view from the eight-language mechanism subset and report the exact severity endpoints.

Table 17: Per-language noisy Recall@100 on the language-balanced MKQA mechanism subset used for script diagnosis; these values are distinct from the full benchmark-wide MKQA column.

Method	Ar	Hi	Bn	Ru	Zh	Ja	Ko	Th
BGE-M3	68.2	64.7	61.3	72.4	74.8	71.2	69.8	63.4
E5-Mistral	67.4	63.8	59.7	71.8	73.9	70.4	68.2	62.1
RAVEL	73.8	71.2	68.7	76.1	77.2	74.8	73.4	70.1

Table 18: Recall@100 across combined-noise severity. The 50% endpoint matches Table 3.

Method	10%	20%	30%	40%	50%
BGE-M3	76.2	68.4	60.1	53.8	49.8
E5-Mistral	77.8	69.1	59.4	52.1	48.3
Translate (ctx.)	74.8	67.2	59.8	54.6	51.2
RAVEL	79.4	73.8	68.2	63.1	58.6

Table 19: Script-family robustness gaps on MKQA; lower is better.

Method	Ar translit.	Cyril.-Lat.	Indic	CJK mix	OCR	Switch	Avg.
Unicode+mContr.	28.4	22.1	31.7	18.3	35.2	24.8	26.8
ByT5	24.1	19.8	27.4	16.2	29.7	21.3	23.1
BGE-M3	21.3	17.4	24.8	14.7	26.1	19.2	20.6
Translate (ctx.)	18.7	15.2	21.3	13.1	28.4	17.8	19.1
RAVEL	12.4	10.8	16.2	9.7	18.3	13.1	13.4

Across the six script/noise columns, RAVEL reduces the average clean-to-noisy gap from BGE-M3's 20.6 points to 13.4. The per-language table also shows positive gains in every evaluated MKQA language rather than a result driven by one script family.

H Derived-Resource Provenance

The bank and transfer tests are separated at the upstream source-family level. Each row identifies the statistics retained for training and the independently sourced representation used for evaluation.

Table 20: Source-family roles and redistribution boundaries for bank construction and source-disjoint evaluation.

Family	Upstream source type	Derived representation	Evaluation boundary
OCR	Tesseract confusion statistics and rescan scripts	Confusion tables and independently rescanned passages	Transfer sources exclude every family used to fit bank statistics; benchmark text remains under original terms.
IME	Public input-method error corpora	Partial-conversion, wrong-mode, and adjacent-key templates	Real-IME instances use disjoint input sources and fixed benchmark labels.
Romanization	Public Arabic–Latin, Cyrillic–Latin, Devanagari–Latin, and Thai–Latin lexicons	Normalized pairs, backoff rules, and false-friend index	Source-disjoint lexicons and instances are excluded from training and bank fitting.
Code-switch	Public multilingual social-media resources	Span-length and language-pair statistics	Evaluation queries come from excluded source families and retain benchmark relevance labels.