AUDITING TRAINING–EVALUATION OVERLAP, NEGATIVE SAMPLING, AND RERANKING IN MULTILINGUAL PATENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-language retrieval systems are usually evaluated against a fixed benchmark, and the resulting numbers are read as evidence about representation quality. We show that in multilingual patent retrieval this reading is fragile: three ordinary design decisions—how negative examples are mined, which overlap definition an audit adopts, and whether a first-stage candidate space is validated—each change what a headline conclusion appears to be. We present CLARITY (Cross-Language Audited Retrieval with Integrity and Transparency), an audited data and evaluation workflow for multilingual patent counterpart retrieval spanning Chinese–English, Japanese–English, and Korean–English (2.692M bilingual records; 23,622 queries; 3,000–4,854 target-language documents per language pair). We contribute (i) a leakage audit that separates two definitions of training–benchmark exposure and shows they differ by 71–837$\times$ across language pairs: 70–76% of evaluation queries appear in the reranker training corpus by query-text match, but only 0.08–1.06% by exact query/positive pair; a control sample of non-benchmark titles matches at 79.7%, and the benchmark loose-exposure rates are 3.5–9.9 pp below this baseline, indicating that query-text overlap reflects corpus vocabulary intersection rather than answer memorization; (ii) a reranking evaluation corrected under the strict definition, where the reranker still improves NDCG@10 by $+0.0249$, $+0.1446$, and $+0.1397$ on the three pairs; and (iii) a repaired hard-negative pipeline: an earlier miner placed the query itself among the negatives in 2.51–9.14% of rows, and after excluding the query, same-component rows, and duplicates, hard negatives beat random negatives on all three language pairs. Baseline measurements for BM25, LaBSE, mE5, BGE-M3, and Qwen3-Embedding show that point-estimate rankings vary across language pairs, and direction asymmetry reaches 0.2619 NDCG@10 for mE5 on Korean–English.

1 INTRODUCTION

Patent search supports prior-art analysis across jurisdictions: relevant technical descriptions can be written in different languages, use specialized vocabulary, and appear in several publications associated with one invention. This makes patents a useful setting for studying multilingual representation learning and retrieval, with a dedicated evaluation history in CLEF-IP (Piroi & Hanbury, 2019; Saad & Nürnberger, 2012) and surveys of the domain's retrieval challenges (Lupu & Hanbury, 2013).

General multilingual encoders are a strong starting point (Feng et al., 2022; Wang et al., 2024; Chen et al., 2024; Qwen Team, 2025; Zhang et al., 2025). Reported performance, however, depends on decisions that are easy to leave implicit: what counts as a query, the language of the candidate documents, the annotated retrieval unit, and where the train/test boundary is drawn. A title query can match the identical title in a same-language document; counting that as cross-language success changes the task. A reranker trained on titles that later appear in the benchmark can memorize rather than generalize, and which of these two effects a paper reports depends on the overlap definition the audit adopts.

CLARITY (Cross-Language Audited Retrieval with Integrity and Transparency) is a data construction and evaluation workflow for multilingual patent counterpart retrieval that treats these decisions

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.

as first-class objects of study. It links corpus construction, documented contrastive training, an explicit exposure audit, and a target-language retrieval benchmark. Our contributions are:

1. **A leakage audit with an explicit definitional choice.** We define and measure two notions of reranker exposure—*loose* (the query text occurs anywhere in the training corpus) and *strict* (the directed query/positive pair occurs)—and show they differ by 71–837$\times$ across language pairs: 70–76% of evaluation queries are loose-exposed against 0.08–1.06% strict-exposed. Inspecting the loose matches resolves the discrepancy: the matched training positives differ from the annotated benchmark counterpart, so such queries teach the model the query text without teaching it the benchmark answer. A control sample of non-benchmark titles matches at 79.7% under the loose test—3.5–9.9 percentage points *above* the benchmark loose-exposure rates (69.8–76.2%), indicating that query-text overlap reflects corpus vocabulary intersection rather than answer memorization.
2. **A leakage-corrected two-stage evaluation.** On the strict-exposure-clean subset the Qwen3-0.6B cross-encoder reranker still improves NDCG@10 by $+0.0249$ (CN–EN), $+0.1446$ (JP–EN), and $+0.1397$ (KR–EN), each with a paired-bootstrap interval excluding zero. Exposed queries gain more, so exposure inflates the raw number but does not create the effect.
3. **A repaired hard-negative pipeline and what the repair changes.** We diagnose a defect in an earlier miner—2.51–9.14% of rows used the query itself as a negative, with up to 12.7% duplicate negatives—and rebuild it to exclude the query, the positive, same-group and same-component rows, and normalized duplicates. Retraining all eighteen cells on the corrected data makes the hard-negative condition consistently *better* than random negatives on all three language pairs, with a mean NDCG@10 advantage of $+0.0064$ (CN–EN), $+0.0199$ (JP–EN), and $+0.0018$ (KR–EN). We report the corrected cross-seed comparison in §4.5; the earlier study's results are not directly comparable due to training–evaluation overlap in the archived data, but the defect rates in Table 3 document the gap between the two mining procedures.
4. **Verified baseline measurements.** BM25, LaBSE, multilingual E5, BGE-M3, and pretrained Qwen3-Embedding are evaluated on all 23,622 queries with paired bootstrap and Holm-adjusted randomization tests; point-estimate rankings vary across language pairs, and direction asymmetry is large for some models.
5. **A text-component-isolated adaptation study.** Training and evaluation do not share normalized title/abstract text, verified at the split level. We state plainly that this is text isolation, not patent-family isolation: the source records carry no family metadata.

2 Related Work

Multilingual representations. Multilingual sentence encoders range from LASER (Artetxe & Schwenk, 2019)—which uses a shared BiLSTM encoder trained on parallel corpora with a max-pooling sentence representation—and LaBSE (Feng et al., 2022)—which adds translation ranking with additive margin softmax to obtain high-quality bidirectional sentence embeddings—to multilingual E5 (Wang et al., 2024), which scales contrastive pretraining over curated multilingual text pairs with consistent "query/passage" instruction prefixes, BGE-M3 (Chen et al., 2024), which unifies dense, sparse (lexical), and multi-vector (ColBERT-style) retrieval in a single model, and Qwen3-Embedding (Qwen Team, 2025; Zhang et al., 2025), which applies inference-time task-specific instructions for retrieval, classification, and clustering. Sentence-BERT (Reimers & Gurevych, 2019) and weakly supervised contrastive pretraining (Wang et al., 2022) motivate reusable text representations, and contrastive objectives of this family trace to InfoNCE (van den Oord et al., 2018). MTEB (Muennighoff et al., 2023) illustrates the need to evaluate representations across tasks. Mr. TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023) evaluate multilingual retrieval across many languages, while BEIR (Thakur et al., 2021) evaluates generalization across heterogeneous retrieval tasks and domains. These benchmarks assume a single canonical document language per query and do not pair source-language queries with target-language documents. In contrast, our benchmark pairs each source-language query with exactly one target-language document per publication, requiring models to bridge languages.

Adaptation and dense retrieval. Domain-adaptive pretraining (Gururangan et al., 2020), dense retrieval (Karpukhin et al., 2020), negative mining strategies such as ANCE (Xiong et al., 2021),

late-interaction scoring in ColBERT (Khattab & Zaharia, 2020), and low-rank adaptation (Hu et al., 2022) are established components for adapting encoders. We apply them to structured bilingual patent views and do not claim that InfoNCE, hard-negative mining, or LoRA is a new objective. Prior patent systems have combined translation and retrieval (Choi, 2009; Saad & Nürnberger, 2012), and patent retrieval has its own survey literature (Lupu & Hanbury, 2013). Importantly, ANCE refreshes negatives from an index that is updated during training; our miner is a one-shot, frozen-encoder procedure, and we name it *static dense hard-negative mining* rather than ANCE-style.

Retrieval and reranking. Two-stage systems pair a first-stage retriever with a more expressive scorer (Nogueira & Cho, 2019; Qu et al., 2021; Ren et al., 2021). Language-model rankers study prompting and scoring choices (Qin et al., 2024; Zhuang et al., 2024); our reranker is a pointwise binary relevance scorer, so we cite those works as background rather than as the method being reproduced.

Train–test overlap and significance testing. Memorized or near-duplicate test instances inflate retrieval and QA metrics in ways that are easy to miss without an explicit audit (Lewis et al., 2021; Elangovan et al., 2021). CLARITY follows this line but operates on publication text, and we add a control sample to distinguish genuine overlap from coincidental matches on generic titles. Existing retrieval benchmarks treat overlap unevenly and, in our reading, do not make it a first-class reported statistic: BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) reuse upstream splits without a pair-level leakage audit, and the closest precedent, Mr. TyDi and MIRACL (Zhang et al., 2021; 2023), separate train/dev/test by language but do not report the strict (query, positive) pair-level overlap we adopt as primary. The pair-level definition is the gap we target: published benchmarks that report only query-text overlap (or none) cannot distinguish corpus vocabulary intersection from answer memorization, which differs by two orders of magnitude in our data. These gaps are not idiosyncratic to patents: whenever a benchmark reuses a large training corpus that also contains the evaluation documents—a common construction in domain adaptation and reranker training—the same two failure modes apply, namely definitional ambiguity (loose vs. strict) and silent data defects (query-as-negative, duplicates) that a raw-equality audit cannot catch. We report paired bootstrap intervals and randomization tests following standard IR practice (Smucker et al., 2007).

3 METHOD

CLARITY links corpus construction, contrastive adaptation, an exposure audit, and a target-language retrieval benchmark (Figure 1).

3.1 LANGUAGE-SPECIFIC RETRIEVAL UNITS

Let q be a title query in language ℓ_q, and let $\mathcal{D}_{\ell_t}$ contain only documents in the target language $\ell_t \neq \ell_q$. A document identifier contains both a publication ID and its language view. For encoder h_θ (parameters θ), the normalized representation f_θ, cosine score s_θ, and top-K set $\mathcal{C}_K$ (retrieving K candidates) are

$$f_\theta(x) = \frac{h_\theta(x)}{\|h_\theta(x)\|_2}, \qquad s_\theta(q,d) = f_\theta(q)^\top f_\theta(d), \qquad \mathcal{C}_K(q) = \mathrm{TopK}_{d \in \mathcal{D}_{\ell_t}} s_\theta(q,d). \tag{1}$$

The paired publication is relevant only in its annotated target-language view. Publication identifiers group the two query directions for statistical inference; they are not used to merge source- and target-language relevance labels.

3.2 BILINGUAL CORPUS AND CONTRASTIVE VIEWS

The archived pipeline extracts source-language and English titles and abstracts from structured patent records. Sampling is stratified by International Patent Classification (IPC) subclass (the four-symbol code, e.g., H01M): for each publication country and each observed IPC subclass, up to 2,000 records are drawn, targeting balanced coverage across technology subclasses rather than proportional sampling. Records are deduplicated across subclasses (a patent may belong to several), and subclasses with fewer than 2,000 patents are taken in full. Table 1 summarizes the collection.

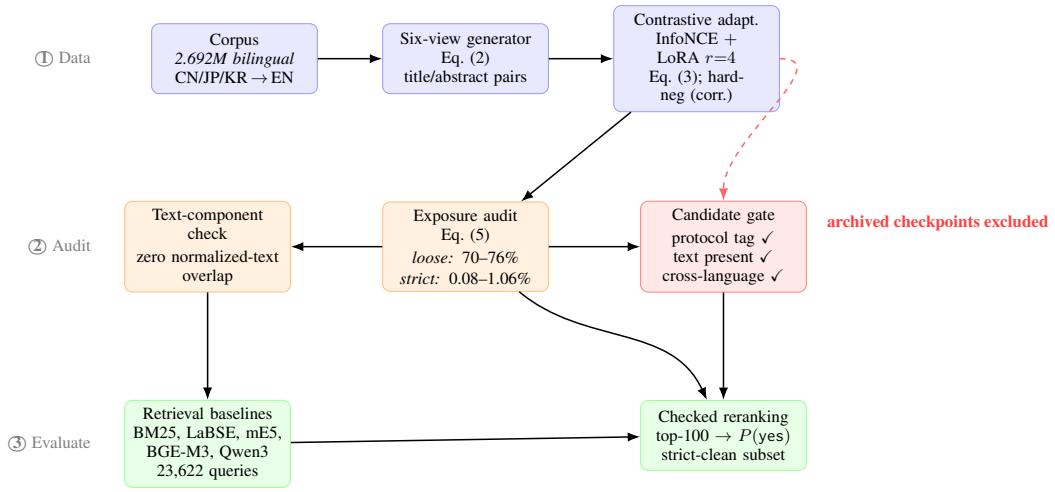


Figure 1: The CLARITY workflow. **Data (purple):** IPC-stratified bilingual corpus → six-view generator (Eq. 2) → contrastive adaptation with corrected hard negatives (Eq. 3). **Audit (orange/red):** exposure audit (Eq. 5) measures loose (70–76%) and strict (0.08–1.06%) exposure—a 71–837× gap—alongside a text-component overlap check and a three-condition candidate gate. **Evaluation (green):** five retrieval baselines benchmarked on 23,622 queries, reranker evaluated on the strict-clean subset. Dashed red arrow: archived checkpoints whose training overlaps the benchmark are excluded.

Table 1: Archived collection statistics. A valid pair has populated source-language and English fields.

Language	Raw files	Valid bilingual records	Yield
CN–EN	1,193,351	1,189,660	99.7%
JP–EN	1,148,065	562,604	49.0%
KR–EN	964,125	939,708	97.5%
Total	3,305,541	2,691,972	81.4%

Given titles (t_s, t_e) and abstracts (a_s, a_e), the six generated positive pairs are

$$(t_s, t_e),\ (t_e, t_s),\ (t_s, a_e),\ (t_e, a_s),\ (t_s, a_s),\ (t_e, a_e). \tag{2}$$

Four views align languages and two connect a title with its abstract in the same language. The archived training runs split rows after view generation and therefore do not implement a publication-disjoint holdout; Table 2 distinguishes potential instance counts from actual saved rows.

3.3 Contrastive Adaptation and Negative Selection

Qwen3-Embedding-0.6B produces 1,024-dimensional representations. With positive d_i^+, negative set $\mathcal{N}_i$, batch size B, and temperature $\tau > 0$, the contrastive loss (van den Oord et al., 2018) is

$$\mathcal{L}_{\text{emb}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s_\theta(q_i, d_i^+)/\tau)}{\exp(s_\theta(q_i, d_i^+)/\tau) + \sum_{d\in\mathcal{N}_i} \exp(s_\theta(q_i, d)/\tau)}. \tag{3}$$

We consider two constructions of $\mathcal{N}_i$. *Random negatives* sample three records other than the current one and take a title or abstract field from each. *Static dense hard negatives* mine the nearest training-pool positives under a frozen encoder, then filter candidates. If we write $g(r)$ and $c(r)$ for the group and component identifiers of row r—where $g(r)$ is the source publication shared by the six views of one record (Eq. 2) and $c(r)$ is the specific title/abstract field combination that identifies a single view—$x_{r'}$ for a text field of candidate row r', and $\eta(\cdot)$ for Unicode NFKC normalization, the admissible set for row r is

$$\mathcal{A}(r) = \left\{ r' \neq r \ :\ g(r') \neq g(r),\ c(r') \neq c(r),\ \eta(x_{r'}) \notin \{\eta(q_r), \eta(d_r^+)\} \right\} \setminus \mathcal{N}_r^{\text{seen}}, \tag{4}$$

Table 2: Archived embedding datasets. Saved rows are counted from the files; potential rows equal six times the valid record count. These archived files overlap evaluation records and are not clean training splits.

Dataset	Valid records	Potential rows	Saved embedding rows
CN–EN (all IPC)	1,189,660	7,137,960	7,137,960
JP–EN (all IPC)	562,604	3,375,624	3,375,624
KR–EN (all IPC)	939,708	5,638,248	5,638,158
Total	2,691,972	16,151,832	16,151,742

where $\mathcal{N}_r^{\text{seen}}$ are candidates already chosen for r, and the emitted row is accepted only if $|\mathcal{N}_r| = 3$ and $\eta(q_r) \notin \{\eta(n) : n \in \mathcal{N}_r\}$. Equation 4 is the operative difference from the earlier miner, whose admissible set checked only $r' \neq r$ and raw inequality with the positive; that gap is what §4.2 measures.

LoRA (Hu et al., 2022) replaces a linear weight W by $W + (\alpha/r)BA$ with $r = 4$, $\alpha = 16$, dropout 0.05, implemented through ms-swift (Zhao et al., 2025). Both the archived and clean-study training runs use the same optimizer (AdamW (Loshchilov & Hutter, 2019)), learning rate 6×10^{-6}, gradient checkpointing, and BF16; they differ in batch size and maximum sequence length, as detailed in Appendix C together with completed steps.

3.4 Exposure Definitions

Let $\mathcal{T}$ be the set of normalized directed (query, positive) strings in a reranker's training corpus and $\mathcal{Q}$ the set of its normalized query strings. For an evaluation query q with positive d^+ and publication $p(q)$,

$$\text{loose}(q) = \mathbf{1}[\eta(q) \in \mathcal{Q}], \qquad \text{strict}(q) = \mathbf{1}[(\eta(q), \eta(d^+)) \in \mathcal{T}]. \tag{5}$$

Here $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 if its argument holds and 0 otherwise, so each definition records whether a single query is exposed. The two definitions answer different questions. *Strict* asks whether the model was ever trained to score exactly this pair, which is the memorization channel that matters for a reranker. *Loose* asks whether the query string appeared at all, which upper-bounds exposure but also fires on generic titles: in a control sample of patent titles that are *not* in the benchmark, 79.7% match the training corpus under the loose test—3.5–9.9 percentage points *higher* than the 69.8–76.2% loose-exposure rates observed for benchmark queries. Inspecting the matches explains why: the matched training positives differ from the annotated benchmark counterpart, so the model saw the query text but never saw it mapped to the evaluation answer. The loose statistic therefore measures corpus vocabulary overlap rather than answer memorization. We report both (Table 10). Figure 2 visualizes this large discrepancy.

4 Experiments

4.1 Evaluation Protocol

The benchmark contains 6,000 CN–EN, 7,914 JP–EN, and 9,708 KR–EN queries, representing 3,000, 3,957, and 4,854 publications with two query directions each. Each target-language corpus contains one document per publication (title and abstract), and a query has one annotated cross-language counterpart, so the task measures paired-record retrieval rather than complete prior-art relevance. Four exact duplicate document-text groups occur in JP; primary results keep the published IDs and single-counterpart labels.

Baselines encode queries and their target corpus afresh with the same model, using a 512-subword maximum length, normalized embeddings, and exact matrix-product search. mE5 (Wang et al., 2024) uses `query:`/`passage:`; Qwen3 (Qwen Team, 2025; Zhang et al., 2025) queries use a fixed patent-retrieval instruction and documents have none; LaBSE (Feng et al., 2022) and BGE-M3 (Chen et al., 2024) add no prompt. BM25 (Robertson & Zaragoza, 2009) provides an untranslated lexical reference ($k_1 = 1.5$, $b = 0.75$, lowercase alphanumeric tokens and individual CJK characters, truncated to 256 tokens) over the same target-language corpus.

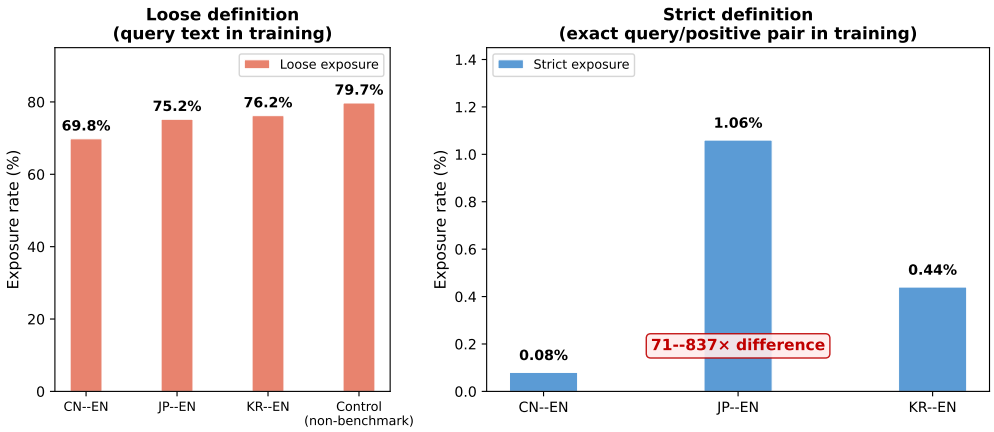


Figure 2: Exposure audit results under two definitions. **Left:** Loose exposure (query text seen in training) affects 70–79.7% of queries, with the control sample (non-benchmark titles) matching above the benchmark rates. **Right:** Strict exposure (exact query/positive pair in training) affects only 0.08–1.06% of benchmark queries—roughly two orders of magnitude less. The loose definition measures vocabulary overlap, not answer memorization.

Table 3: Defects in the earlier static hard-negative files (60,000-row sample per language) versus the rebuilt miner over all rows. The rebuilt files are the data used for every hard-negative result in this paper.

Language	Query=negative (old)	Duplicate negatives (old)	Query=negative (new)	Duplicate negatives (new)
CN–EN	2.51%	2.08%	0%	0%
JP–EN	9.14%	12.74%	0%	0%
KR–EN	4.35%	5.85%	0%	0%

We report NDCG@10, Recall@10, and AP@10. With one relevant target, AP@10 and MRR@10 coincide, and for rank r_q of the relevant document ($\mathbf{1}[\cdot]$ as defined under Eq. 5),

$$\text{NDCG@10} = \frac{\mathbf{1}[r_q \le 10]}{\log_2(r_q + 1)}, \qquad \text{Recall@10} = \mathbf{1}[r_q \le 10]. \tag{6}$$

Uncertainty uses 10,000 paired bootstrap draws over publications, keeping both query directions together; a two-sided Monte Carlo sign randomization test also operates on publications with $p = (b+1)/(10{,}000+1)$, and Holm correction covers the nine Qwen3-versus-neural-baseline language comparisons.

4.2 The Hard-Negative Defect and Its Consequences

An earlier static miner produced three files that were used for the hard-negative condition of the adaptation study. Auditing a 60,000-row sample per language against Equation 4 shows the defects in Table 3: a non-trivial fraction of rows carried the query itself as one of their negatives, and a comparable fraction carried repeated negatives. The accompanying audit artifact reported `false_negatives: 0` because it compared only raw string equality against the positive and never checked the query, group membership, or normalized duplicates.

The rebuilt miner emits 341,940 / 341,880 / 342,312 rows for CN/JP/KR and was audited over all 1,026,132 rows, again recovering zero query-equals-negative, zero duplicate-negative, and zero same-component rows. We then retrained every hard-negative checkpoint on this corrected data under the identical budget. We do not compare old-defective against new-corrected hard negatives on the full benchmark, because the archived checkpoints share training text with the evaluation partition; the corrected comparison is in §4.5.

Table 4: Full-scale reranking on the strict-exposure-clean subset, where no directed (query, positive) pair from the query appears in the reranker training corpus. All differences have paired-bootstrap intervals excluding zero and randomization $p < 10^{-4}$.

Language pair	Clean n	Embedding	+ Reranker	Δ	95% CI
CN–EN	5,995	0.9687	**0.9936**	+0.0249	[+0.0213, +0.0286]
JP–EN	7,830	0.7581	**0.9027**	+0.1446	[+0.1377, +0.1516]
KR–EN	9,665	0.7712	**0.9109**	+0.1397	[+0.1331, +0.1465]

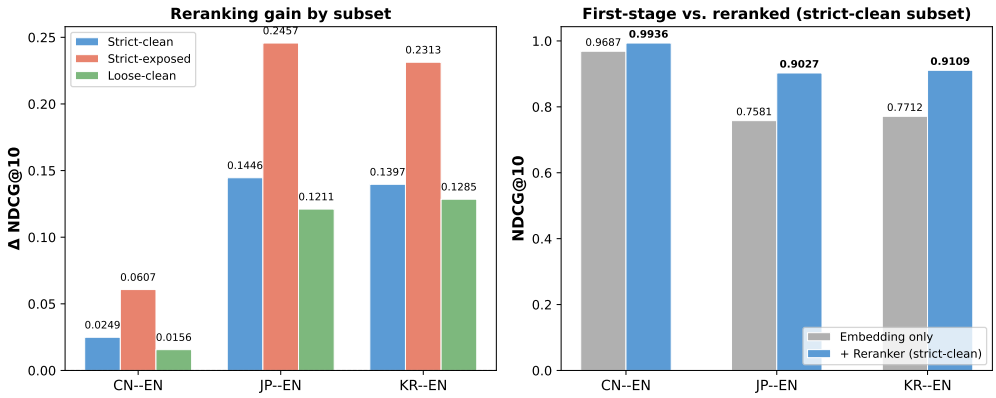


Figure 3: Reranking evaluation under strict-exposure correction. **Left:** ΔNDCG@10 for three exposure subsets: strict-clean, strict-exposed, and loose-clean. Exposed queries gain more, but the clean subset still shows significant improvement on every language pair. **Right:** First-stage vs. reranked NDCG@10 on the strict-clean subset. The reranker consistently improves over the embedding-only baseline.

4.3 Leak-Corrected Reranking

A Qwen3-0.6B cross-encoder reranker, previously trained on an all-language reranking corpus (5M shuffled examples with balanced CN/JP/KR language ratios, LoRA rank 4, learning rate 6×10^{-6}, 512 maximum length, binary relevance scoring with yes/no logit softmax), is applied to the top-100 Qwen3-Embedding-0.6B candidates under the strict benchmark protocol, with validation that every candidate text is present and that no candidate shares the query's own language view. Full training and scoring details are in Appendix E.

An audit of that historical reranker training corpus against the benchmark reveals that, under the strict definition of Equation 5, only 0.08% (CN–EN), 1.06% (JP–EN), and 0.44% (KR–EN) of evaluation queries have an exact directed (query, positive) pair present in training. This overlap arises because the archived pipeline lacked an explicit exposure check—the gap CLARITY fills. Rather than discarding the reranker, we recompute every statistic on the strict-clean subset and verify whether the reranking conclusion survives the correction.

Table 4 reports the result: the gain survives the correction on all three pairs, and the clean subset is 99.9%, 98.9%, and 99.6% of the queries respectively.

Exposure inflates the uncorrected number: the four-way split (Table 10) shows exposed queries gain $+0.0607$, $+0.2457$, $+0.2313$ (strict) vs. $+0.0249$, $+0.1446$, $+0.1397$ for clean; clean-loose CN–EN still gains $+0.0156$. The CN–EN strict-exposed gain rests on only n=5 queries and should not be over-interpreted; the JP–EN and KR–EN strict-exposed subsets (n=84 and n=43) are the substantive cases. Removing the most exposed queries lowers the estimated gain but never to zero. Exposed queries also have lower initial scores (e.g., JP strict-exposed 0.3356 vs. clean 0.7581), so the larger gain may reflect difficulty rather than exposure alone. Figure 3 visualizes these gains.

Table 5: NDCG@10 under target-language-only retrieval. Both directions are averaged per language pair. Bold: best per column (neural models only).

Model	CN–EN	JP–EN	KR–EN
BM25 (untranslated)	0.0497	0.0018	0.0372
LaBSE	0.6287	0.3860	0.3708
mE5-large	0.8529	**0.7572**	0.6743
BGE-M3	0.8636	0.6314	0.7597
Qwen3-Embedding-0.6B	**0.9684**	0.7536	**0.7690**

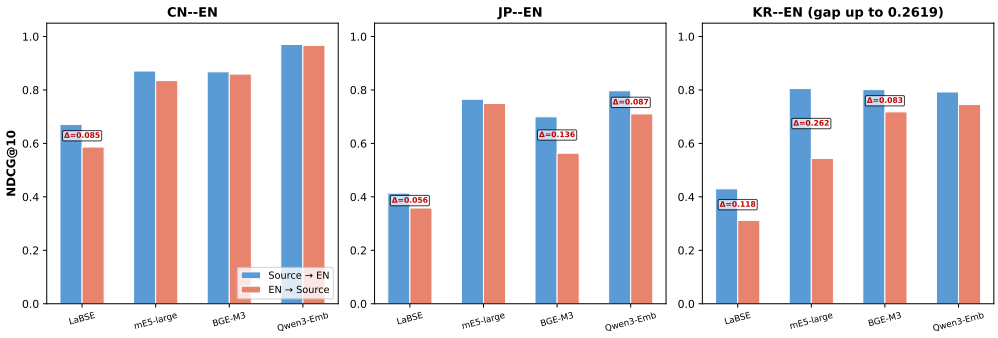


Figure 4: Direction asymmetry across encoders: source-language→English vs. English→source-language NDCG@10. Blue bars show the source-to-English direction; red bars show the reverse. Large gaps (annotated in red) are most pronounced in KR–EN, where mE5 exhibits a 0.2619 gap.

4.4 Verified Baselines

Table 5 reports the untranslated BM25 reference and four pretrained encoders. None of the neural models is adapted on our training files in this evaluation.

Qwen3 obtains the highest point estimate on CN–EN and KR–EN. The CN advantage over BGE-M3 is 0.1048 NDCG@10 (95% paired CI [0.0967, 0.1132]); the KR advantage is 0.0093 ([0.0018, 0.0170], Holm-adjusted $p = 0.0344$). On JP–EN, mE5 is numerically higher and the Qwen3-minus-mE5 difference is -0.0036 ([-0.0123, 0.0050], adjusted $p = 0.4070$). The data therefore do not support a universal ordering of the strongest multilingual encoders. BM25 obtains NDCG@10 of 0.0497 (CN–EN), 0.0018 (JP–EN), and 0.0372 (KR–EN); its non-zero CN–EN score is attributable to shared Latin-alphabet tokens rather than cross-language semantic matching, as a hit-query analysis confirms; full metrics appear in Appendix A.

Direction-specific measurements (Appendix F) reveal further differences. On KR–EN, mE5 scores 0.8052 from Korean to English and 0.5434 in reverse, a gap of 0.2619; Qwen3 scores 0.7925 and 0.7454. Averaging directions conceals this variation. Figure 4 visualizes the direction gap across all models and language pairs. These measurements describe retrieval asymmetry under the present construction and do not isolate its linguistic cause.

4.5 Adaptation under Text Isolation

The archived adaptation corpus overlaps the benchmark, so its checkpoints cannot support held-out claims. We build a split whose train and dev partitions share no normalized title/abstract text (Table 6); the grouping is over normalized shared-text connected components, *not* patent families, and the source records carry no family metadata, so family disjointness is unknown. Since the adaptation study evaluates only on its dev partition, the absolute NDCG@10 values in Table 7 are *not comparable* to the pretrained-encoder scores in Table 5; all comparisons here are between the two negative-sampling conditions within the same evaluation subset.

Within this split we train eighteen LoRA checkpoints: three seeds $\{13, 42, 2027\}$ crossed with two negative-sampling conditions, for each language. Every run uses a fixed 512-update budget with an

Table 6: Text-isolated training splits for the adaptation study. Isolation is verified over normalized title/abstract text; it is not patent-family isolation.

Language	Train groups	Dev groups	Seeds	Status
CN–EN	56,990	3,010	13, 42, 2027	trained
JP–EN	56,980	3,020	13, 42, 2027	trained
KR–EN	57,052	2,948	13, 42, 2027	trained

Table 7: Adaptation results across three seeds on the text-isolated split. Each cell aggregates three independently trained LoRA checkpoints; entries are mean ± standard deviation over seeds. The comparison of interest is between the two negative-sampling conditions within each row.

Language pair	Negative sampling	Seeds	NDCG@10	Recall@10
CN–EN	Random	13, 42, 2027	0.9792 ± 0.0005	0.9952 ± 0.0003
CN–EN	Hard (corrected)	13, 42, 2027	$\mathbf{0.9856 \pm 0.0004}$	$\mathbf{0.9958 \pm 0.0002}$
JP–EN	Random	13, 42, 2027	0.8270 ± 0.0008	0.9142 ± 0.0007
JP–EN	Hard (corrected)	13, 42, 2027	$\mathbf{0.8469 \pm 0.0012}$	$\mathbf{0.9259 \pm 0.0014}$
KR–EN	Random	13, 42, 2027	0.8187 ± 0.0011	0.9032 ± 0.0007
KR–EN	Hard (corrected)	13, 42, 2027	$\mathbf{0.8205 \pm 0.0009}$	$\mathbf{0.9045 \pm 0.0010}$

effective batch size of 256 examples per update (per-device batch size 32 × gradient accumulation 8), roughly 0.38 passes over the training partition; we state the update budget rather than an epoch count because the runs are update-limited. Table 7 reports the result.

Seed-to-seed variance is small, so the procedure is stable. With the corrected miner, hard negatives beat random negatives on all three pairs, with a modest but uniform effect: $+0.0064$ (CN–EN), $+0.0199$ (JP–EN), $+0.0018$ (KR–EN). The KR–EN gain is small relative to seed variability ($\sigma \approx 0.001$), so we extend it to eight seeds; the gain stays positive ($+0.0029$, ahead on each seed)—nonparametric evidence of a small effect, not a significance test (Appendix I). This is a cautionary result: a quality-control gap in a few percent of rows obscured the comparison.

4.6 Ablations

We test $k \in \{1, 3, 5, 10\}$ and $\tau \in \{0.05, 0.07, 0.1\}$ on the text-isolated split with seed $\{13, 42, 2027\}$ and the same budget; $k = 3$ and $\tau = 0.1$ reproduce Table 7. Both have a small effect on the main (hard-negative) condition: across the tested k and τ ranges the hard-negative score moves by at most $\sim$0.02 NDCG@10. The hard-over-random advantage is specific to the main setting τ=0.1: at lower τ the random-negative baseline scores higher than hard on JP–EN and KR–EN, so the hard-over-random comparison is not uniform across τ. We retain $k = 3$ as a compute-efficient middle value and τ=0.1 as the main setting. Full values are in Appendix H.

5 Discussion and Conclusion

CLARITY treats audit definitions as findings rather than footnotes. (1) Model ordering is not universal: rankings vary across pairs; direction-specific reporting exposes gaps up to 0.2619 NDCG@10 (mE5, KR–EN), and Qwen3-Embedding is strongest on CN–EN and KR–EN, not significantly different from mE5 on JP–EN. (2) Reranking gains are real under both exposure definitions but size is definition-dependent: the strict pair-level one matters for a reranker, while the query-text criterion is insensitive to memorization; report both, pair-level primary. (3) Quality-control gaps yield uninterpretable comparisons: an earlier miner carried defects in 2.51–9.14% of rows, undetected by raw-string equality; after correction, hard negatives beat random negatives on every pair ($+0.0064$, $+0.0199$, $+0.0018$), so audits must check the query, groups, and normalized duplicates. These findings are not patent-specific; they recur whenever retrieval evaluation reuses a corpus containing the evaluation documents. **Limitations.** The benchmark measures single-publication retrieval, not production indexing; family and pretraining-corpus membership are unavailable (Lewis et al., 2021; Elangovan et al., 2021), and our exact-string audit misses semantic and pretraining-data overlap, so strict-exposure rates are lower bounds (Appendix B).

AI USE STATEMENT

AI assistants were used for language polishing and proofreading of the manuscript. All experiments, data, and reported results are conducted by the authors.

REFERENCES

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://doi.org/10.1162/tacl_a_00288.

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/.

YooChan Choi. Korean to english patent automatic translation (k2e-pat) and cross lingual retrieval on kipris. *World Patent Information*, 31(2):135–136, 2009. doi: 10.1016/j.wpi.2008.09.005. URL https://doi.org/10.1016/j.wpi.2008.09.005.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1325–1335, 2021. doi: 10.18653/v1/2021.eacl-main.113. URL https://doi.org/10.18653/v1/2021.eacl-main.113.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, 2022. doi: 10.18653/v1/2022.acl-long.62. URL https://doi.org/10.18653/v1/2022.acl-long.62.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://doi.org/10.18653/v1/2020.acl-main.740.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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://doi.org/10.18653/v1/2020.emnlp-main.550.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1000–1008, 2021. doi: 10.18653/v1/2021.eacl-main.86. URL https://doi.org/10.18653/v1/2021.eacl-main.86.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mihai Lupu and Allan Hanbury. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7(1):1–97, 2013. doi: 10.1561/1500000027. URL https://doi.org/10.1561/1500000027.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426.

Niklas Muennighoff, Nouamane Tazi, Loic 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://doi.org/10.18653/v1/2023.eacl-main.148.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT, 2019. URL https://arxiv.org/abs/1901.04085. arXiv:1901.04085.

Florina Piroi and Allan Hanbury. Multilingual patent text retrieval evaluation: Clef-ip. In *Information Retrieval Evaluation in a Changing World*, pp. 365–387, 2019. doi: 10.1007/978-3-030-22948-1_15. URL https://doi.org/10.1007/978-3-030-22948-1_15.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024. doi: 10.18653/v1/2024.findings-naacl.97. URL https://doi.org/10.18653/v1/2024.findings-naacl.97.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, 2021. doi: 10.18653/v1/2021.naacl-main.466. URL https://doi.org/10.18653/v1/2021.naacl-main.466.

Qwen Team. Qwen3-Embedding-0.6B model card, 2025. URL https://huggingface.co/Qwen/Qwen3-Embedding-0.6B. Hugging Face model card, accessed 2026-09-12.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://doi.org/10.18653/v1/D19-1410.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, 2021. doi: 10.18653/v1/2021.emnlp-main.224. URL https://doi.org/10.18653/v1/2021.emnlp-main.224.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Farag Saad and Andreas Nürnberger. Overview of prior-art cross-lingual information retrieval approaches. *World Patent Information*, 34(4):304–314, 2012. doi: 10.1016/j.wpi.2012.08.013. URL https://doi.org/10.1016/j.wpi.2012.08.013.

Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp. 623–632, 2007. doi: 10.1145/1321440.1321528. URL https://doi.org/10.1145/1321440.1321528.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

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

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*, 2024. doi: 10.48550/arXiv.2402.05672. URL https://arxiv.org/abs/2402.05672.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.00808.

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://doi.org/10.18653/v1/2021.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://doi.org/10.1162/tacl_a_00595.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025. doi: 10.1609/aaai.v39i28.35383. URL https://ojs.aaai.org/index.php/AAAI/article/view/35383. Software package ms-swift: https://github.com/modelscope/ms-swift.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. 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. 358–370, 2024. doi: 10.18653/v1/2024.naacl-short.31. URL https://doi.org/10.18653/v1/2024.naacl-short.31.

A Detailed Baseline Results

Intervals resample publications. They quantify sampling uncertainty conditional on the fixed trained checkpoints; they do not quantify training-seed variance or family-level dependence. Table 8 reports the full per-model metrics with 95% paired bootstrap intervals for the neural models; BM25, as a parameter-free lexical reference, is reported without an interval.

BM25 CN–EN hit analysis. Of the 6,000 CN–EN queries, 366 (6.1% Recall@10) are hit by BM25. Of these hit queries, 86.6% contain English words, 51.6% contain patent identifiers or special codes

Table 8: Baseline details; AP@10 equals MRR@10 under the single-target labels. BM25 is an untranslated lexical reference; its very low Recall@10 and AP@10 (e.g. 0.0013 AP@10 on JP–EN) reflect the near-absence of cross-language lexical overlap rather than a model limitation, and a hit-query analysis explaining the non-zero CN–EN score appears below.

Language	Model	NDCG@10 [95% CI]	Recall@10	AP@10
CN–EN	BM25	0.0497	0.0610	0.0462
CN–EN	LaBSE	0.6287 [0.6156, 0.6416]	0.7868	0.5786
CN–EN	mE5-large	0.8529 [0.8438, 0.8615]	0.9368	0.8258
CN–EN	BGE-M3	0.8636 [0.8544, 0.8724]	0.9375	0.8399
CN–EN	Qwen3-Embedding-0.6B	**0.9684** [0.9641, 0.9725]	**0.9923**	**0.9605**
JP–EN	BM25	0.0018	0.0033	0.0013
JP–EN	LaBSE	0.3860 [0.3743, 0.3979]	0.5215	0.3437
JP–EN	mE5-large	**0.7572** [0.7477, 0.7669]	**0.8576**	**0.7252**
JP–EN	BGE-M3	0.6314 [0.6195, 0.6432]	0.7354	0.5985
JP–EN	Qwen3-Embedding-0.6B	0.7536 [0.7439, 0.7633]	0.8534	0.7217
KR–EN	BM25	0.0372	0.0509	0.0329
KR–EN	LaBSE	0.3708 [0.3607, 0.3811]	0.5161	0.3258
KR–EN	mE5-large	0.6743 [0.6649, 0.6839]	0.7765	0.6420
KR–EN	BGE-M3	0.7597 [0.7504, 0.7688]	0.8503	0.7308
KR–EN	Qwen3-Embedding-0.6B	**0.7690** [0.7600, 0.7778]	**0.8644**	**0.7384**

(CAS, IPC), and 26.0% contain digits, compared to 50.0%, 1.8%, and 1.2% respectively among missed queries. The non-zero CN–EN BM25 score is therefore driven by shared Latin-alphabet tokens (chemical formulas, patent codes, numerals) rather than cross-language semantic matching. The near-zero JP–EN score (0.0018) reflects the minimal lexical overlap between untranslated Japanese and English under character-level tokenization, and does not imply that lexical methods are inherently unsuitable with appropriate translation or tokenization.

B Scope of the Exposure Audit

Our audit covers the reranker training corpus that we can enumerate, using NFKC/lowercase/whitespace-normalized exact string matching. It does not detect semantic-level overlap (near-duplicate texts, paraphrases, or patent family members with distinct wording), query–answer pairs that co-occur in the training corpus with different field combinations, or documents that appeared in the base model's pretraining data. The 0.08–1.06% strict-exposure rates reported in §4.3 should therefore be understood as lower bounds under the specific matching criterion defined in Equation 5, not as comprehensive leakage-free guarantees. Future work should extend these audits to semantic similarity thresholds, patent-family-level overlap, and pretraining-data membership inference.

C Training Configuration

The archived all-IPC adaptation used the saved embedding files in Table 2, a row-level validation fraction of 10^{-4}, and seed 42. The optimizer is AdamW (Loshchilov & Hutter, 2019) with learning rate 6×10^{-6}, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.95)$, cosine decay, no warmup, and gradient clipping at 1.0, in BF16 with gradient checkpointing and maximum sequence length 4,096. These configurations are training provenance, and overlap with evaluation records precludes interpreting the resulting scores as held-out gains.

The corrected clean-study runs use maximum length 512, per-device batch 32, gradient accumulation 8, LoRA rank 4, alpha 16, dropout 0.05, all-linear target modules, learning rate 6×10^{-6}, and a fixed 512-update budget. Each run's completed step count and final loss are recorded in its `trainer_state.json`; all eighteen cells report 512/512 updates.

D Encoder Interfaces

The fixed checkpoints are `sentence-transformers/LaBSE`, `intfloat/multilingual-e5-large`, `BAAI/bge-m3`, and `Qwen/Qwen3-Embedding-0.6B`, with embedding dimensions 768, 1,024, 1,024, and 1,024. All run in BF16 through SentenceTransformers, with normalized embeddings converted to float32 for matrix-product ranking and top-100 rankings retained. The 512-token cap includes prompts and special tokens. Qwen3 prefixes each query with `Instruct:` followed by "Given a patent title, retrieve its corresponding patent document in the target language.", a newline, and `Query:`; its documents have no prefix.

E Reranking Pipeline and Exposure Audit

The reranker is a merged LoRA checkpoint of `Qwen/Qwen3-0.6B` (the same 1,024-hidden-size, 28-layer causal language model as the embedding encoder), loaded with `local_files_only`. A pair is formatted as a chat template with a fixed system instruction, followed by a user turn containing `<Instruct>`, `<Query>`, and `<Document>` fields; the score is $P(\texttt{yes})$ over the `yes`/`no` logits. Before scoring, the pipeline asserts that the input file carries the `target_language_exact_qrels_v2` protocol tag, raises on any missing or empty candidate text, and raises if a candidate shares the query's own language-view suffix.

The exposure audit streams the reranker training corpus once, normalizing every query and positive string, and intersects it with the benchmark's corpus and query strings. The control sample draws 2,000 patent titles from the archived corpora that are absent from the benchmark, and measures the same intersection; the resulting 79.7% loose hit rate is why we treat strict matching as the primary definition. All audit code and the resulting partitions are released with the paper.

F Direction-Specific Results

Each direction has 3,000, 3,957, or 4,854 queries for CN, JP, or KR. Source language refers to Chinese, Japanese, or Korean; each target corpus has one document per publication. Table 9 reports NDCG@10 by query direction for the four pretrained encoders.

Table 9: NDCG@10 by query direction for the four pretrained encoders.

Language	Model	Source → EN	EN → source
CN–EN	LaBSE	**0.6710**	0.5864
CN–EN	mE5-large	**0.8708**	0.8350
CN–EN	BGE-M3	**0.8679**	0.8594
CN–EN	Qwen3-Embedding-0.6B	**0.9702**	0.9667
JP–EN	LaBSE	**0.4138**	0.3582
JP–EN	mE5-large	**0.7652**	0.7493
JP–EN	BGE-M3	**0.6995**	0.5633
JP–EN	Qwen3-Embedding-0.6B	**0.7969**	0.7103
KR–EN	LaBSE	**0.4299**	0.3117
KR–EN	mE5-large	**0.8052**	0.5434
KR–EN	BGE-M3	**0.8013**	0.7180
KR–EN	Qwen3-Embedding-0.6B	**0.7925**	0.7454

G Reranking Exposure Sensitivity

Table 10 reports the four-way split of every language pair by exposure definition. For each pair, the strict-clean subset excludes queries whose directed (query, positive) pair appears in the reranker training corpus, while the loose-clean subset additionally excludes any query whose text appears anywhere in training; the exposed subsets contain the remaining queries. The loose control rate on non-benchmark titles is 79.7%, making loose-clean the more conservative subset. Reranking gains are positive on every subset, but exposed queries gain more, so the strict-clean column is the one that isolates the generalization effect.

Table 10: Sensitivity of the reranking conclusion to the exposure definition. Strict excludes directed training pairs; loose excludes any query text seen in training. The loose control rate on non-benchmark titles is 79.7%, so loose-clean is a conservative subset.

Language	Subset	n	Embedding	+ Reranker	Δ
CN–EN	Clean (strict)	5,995	0.9687	**0.9936**	+0.0249
CN–EN	Exposed (strict)	5	0.6602	**0.7209**	+0.0607
CN–EN	Clean (loose)	1,813	0.9801	**0.9958**	+0.0156
CN–EN	Exposed (loose)	4,187	0.9634	**0.9923**	+0.0289
JP–EN	Clean (strict)	7,830	0.7581	**0.9027**	+0.1446
JP–EN	Exposed (strict)	84	0.3356	**0.5813**	+0.2457
JP–EN	Clean (loose)	1,964	0.8231	**0.9442**	+0.1211
JP–EN	Exposed (loose)	5,950	0.7307	**0.8845**	+0.1538
KR–EN	Clean (strict)	9,665	0.7712	**0.9109**	+0.1397
KR–EN	Exposed (strict)	43	0.2704	**0.5017**	+0.2313
KR–EN	Clean (loose)	2,311	0.8015	**0.9300**	+0.1285
KR–EN	Exposed (loose)	7,397	0.7588	**0.9026**	+0.1438

H Hyperparameter Ablation Tables

Tables 11 and 12 report the full sensitivity of the adaptation procedure to the hard-negative count k and the InfoNCE temperature τ on the text-isolated split. Across the tested k and τ ranges the hard-negative score moves by at most $\sim$0.02 NDCG@10. The hard-over-random advantage reported in Table 7 holds at the main setting τ=0.1, but is not uniform across τ: at $\tau \in \{0.05, 0.07\}$ the random-negative baseline outscores hard on JP–EN and KR–EN. The main results should therefore be read as the hard-over-random comparison at τ=0.1, which is the setting we adopt. The $k = 3$ and $\tau = 0.1$ columns reproduce the main hard-negative condition of Table 7.

Table 11: NDCG@10 sensitivity to hard-negative count k on the text-isolated split. Each cell is mean $\pm$ standard deviation over three seeds. The k=3 row reproduces the main hard-negative condition from Table 7.

Language	k=1	k=3 (main)	k=5	k=10
CN–EN	0.9852 ± 0.0005	0.9856 ± 0.0004	0.9855 ± 0.0007	0.9855 ± 0.0007
JP–EN	0.8347 ± 0.0014	0.8469 ± 0.0012	0.8494 ± 0.0012	0.8556 ± 0.0011
KR–EN	0.8151 ± 0.0008	0.8205 ± 0.0009	0.8230 ± 0.0003	0.8250 ± 0.0004

Table 12: NDCG@10 sensitivity to InfoNCE temperature τ on the text-isolated split. Each cell is mean $\pm$ standard deviation over three seeds. The $\tau = 0.1$ cells reproduce the main results from Table 7.

Language	Condition	τ=0.05	τ=0.07	τ=0.1 (main)
CN–EN	Random	0.9849 ± 0.0004	0.9842 ± 0.0004	0.9792 ± 0.0005
CN–EN	Hard	0.9869 ± 0.0001	0.9867 ± 0.0004	0.9856 ± 0.0004
JP–EN	Random	0.8556 ± 0.0005	0.8516 ± 0.0007	0.8270 ± 0.0008
JP–EN	Hard	0.8340 ± 0.0019	0.8441 ± 0.0006	0.8469 ± 0.0012
KR–EN	Random	0.8511 ± 0.0007	0.8411 ± 0.0007	0.8187 ± 0.0011
KR–EN	Hard	0.8242 ± 0.0015	0.8245 ± 0.0010	0.8205 ± 0.0009

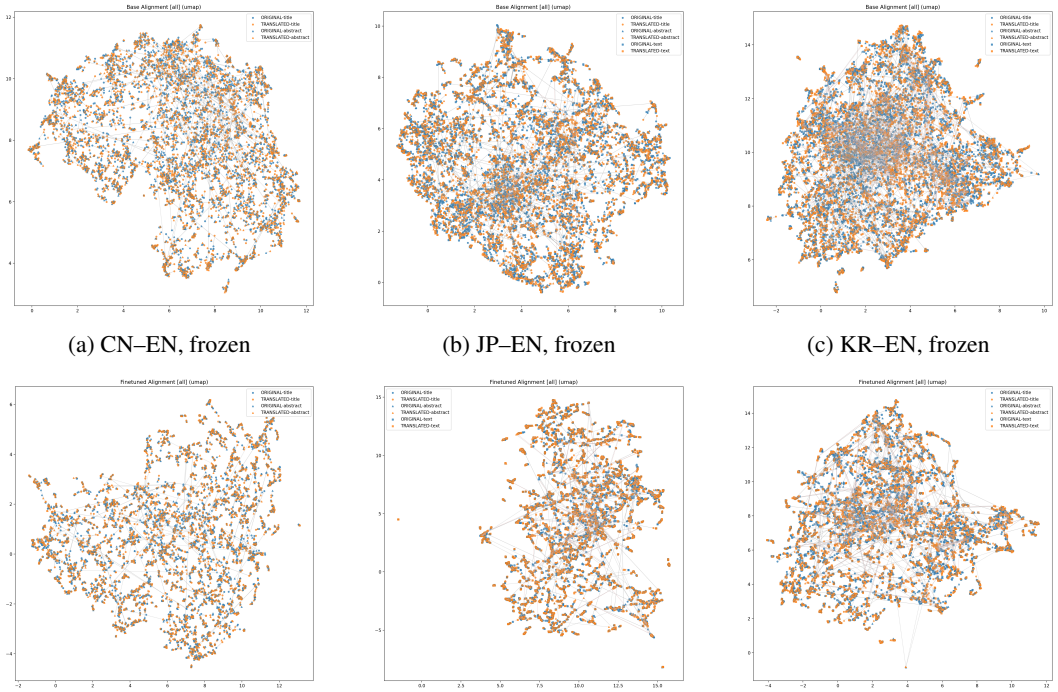
I Hard-vs-Random: Statistical Note

We do not report a hard-vs-random paired-bootstrap interval because the two conditions are not paired at the query level across independent training runs: each seed trains a separate hard-negative

and random-negative checkpoint, so the per-query outcomes are not tied. Instead we rely on cross-seed consistency. Over three seeds $\{13, 42, 2027\}$ the hard-negative mean exceeds the random-negative mean for all three language pairs ($+0.0064$, $+0.0199$, $+0.0018$ for CN/JP/KR). Because the KR–EN gain is small relative to seed variability ($\sigma \approx 0.001$), we extend KR–EN only to eight seeds and the KR–EN gain stays positive ($+0.0029$) and does not change sign across seeds. This is non-parametric evidence of a small but real effect rather than a formal significance test; the KR–EN result should be read as "within seed variability, consistent direction."

J EMBEDDING VISUALIZATION

To qualitatively assess the effect of contrastive adaptation, we project the normalized embeddings of 2,000 patent title–abstract pairs per language pair into two dimensions using UMAP (McInnes et al., 2018) and t-SNE (van der Maaten & Hinton, 2008). For each pair, the source-language title/abstract and its English counterpart are encoded with the frozen Qwen3-Embedding-0.6B and the adapted LoRA checkpoint (hard-negative condition, seed 42). Figure 5 places the frozen and adapted checkpoints side by side for each language pair: under the frozen encoder the source (blue) and English (orange) clouds overlap only partially, whereas after adaptation the two clouds are visibly closer and more compact. The frozen and adapted panels are fitted with separate UMAP reducers, so the comparison is qualitative within each panel rather than a quantitative distance comparison.



(a) CN–EN, frozen (b) JP–EN, frozen (c) KR–EN, frozen

(d) CN–EN, adapted (e) JP–EN, adapted (f) KR–EN, adapted

Figure 5: Frozen vs. adapted UMAP comparison (2,000 samples per language pair). Top row: frozen Qwen3-Embedding-0.6B; bottom row: adapted LoRA checkpoint (hard-negative, seed 42). Compare top-to-bottom within each column (same language pair): under the frozen encoder the source (blue) and English (orange) clouds are only partially overlapping, whereas after adaptation they are visibly closer and more compact, matching the quantitative gain in Table 7.

All 36 visualization figures (3 languages $\times$ 2 models $\times$ 2 reduction methods $\times$ 3 view modes) are included in the open-source release.