EMBEDFLOW: UPGRADING LEGACY EMBEDDINGS WITHOUT FULL UPFRONT RE-EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Upgrading an embedding model usually means rebuilding the document index. This is a natural consequence of independently trained models living in different vector spaces: embeddings from the new model generally cannot be searched directly against vectors produced by the old one. But this does not necessarily make the old index useless. We ask whether a legacy retriever can still return enough useful candidates for a new model to recover its retrieval quality without first re-embedding the full corpus. We call this property *candidate compatibility*.

We study candidate compatibility across 63 source–target migrations. At $K = 50$, 42/63 settings are within 0.01 nDCG@10 of native target retrieval. The same pattern appears in a frozen BRIGHT evaluation and remains stable in a controlled Natural Questions study from 100K to one million documents. At 1M documents, Qwen3-4B→8B is effectively lossless with only 50 source candidates. Qwen3-0.6B→8B and MiniLM→8B require deeper pools, reaching the same stringent near-target criterion at $K = 200$ and $K = 500$, respectively.

A practical challenge is deciding whether a migration is compatible without first building the target index. To address this, we introduce a frozen residual candidate-tail diagnostic that uses neither relevance labels nor native target retrieval. Across 28 frozen holdout cells, it makes 17 `SAFE` predictions with zero observed false-safes, and it recovers the same migration ordering in a subsequent blind million-document evaluation.

Finally, we study what happens when the legacy index itself is approximate. We show that deployed migration error separates into two terms: an intrinsic source–target compatibility gap and an additional penalty from ANN search. Our million-document ANN experiments match this decomposition. Together, these results suggest that embedding migration need not be treated as an all-or-nothing re-indexing operation. For compatible model transitions, a legacy index can remain useful while the new model is introduced progressively.

1 INTRODUCTION

Embedding models sit at the core of modern search and retrieval-augmented generation systems (Lewis et al., 2020). Upgrading them, however, is often surprisingly expensive. Independently trained encoders generally occupy different vector spaces, so query embeddings from a new model cannot be searched directly against document vectors produced by an older one. In practice, this usually means re-embedding the corpus and rebuilding the index before the new model can be deployed.

Most prior work attacks this problem by making old and new representations interoperable. This includes backward-compatible training, partial backfilling, and post-hoc mappings between embedding spaces (Shen et al., 2020; Ramanujan et al., 2022; Zhang et al., 2022; Hu et al., 2022; Jaeckle et al., 2023; Cai et al., 2023; Conneau et al., 2018; Maystre et al., 2025; Yang et al., 2025; Vejendla, 2025; Maekawa et al., 2026). We study a different question: *what if direct interoperability is not actually necessary?*

Suppose the legacy retriever still returns many of the documents that the new model would want to see. In that case, the old index may remain useful even though its vectors are geometrically incom-

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.

patible with the target model. The legacy model can generate candidates, while the target model scores those documents in its own embedding space. We call this property *candidate compatibility*.

The retrieve–then–rerank pattern itself is not new; multi-stage retrieval is well established (Karpukhin et al., 2020; Nogueira & Cho, 2019; Khattab & Zaharia, 2020). What is different here is the migration setting. The first-stage retriever is not a component chosen for the new system: it is a historical index that is already deployed, and the target model was trained independently of it. This raises a different set of questions. How much target quality can the legacy retriever preserve? How deep does the source candidate pool need to be? Does the relationship remain stable as the corpus grows? Can compatibility be estimated before building a full target index? And how much does approximate nearest-neighbor search change the answer?

We study these questions across increasingly demanding settings. In a 63-cell development sweep, 42/63 migrations are within 0.01 nDCG@10 of native target retrieval at $K = 50$, with 25/45 still meeting the same margin after removing the two most saturated Nano tasks. A frozen BRIGHT evaluation reproduces the same migration hierarchy, and the ordering remains stable on nested Natural Questions corpora from 100K to one million documents. At 1M documents, Qwen3-4B→8B is effectively lossless at $K = 50$, while Qwen3-0.6B→8B and MiniLM→8B require deeper candidate sets.

We also study whether compatibility can be estimated before a native target index exists. A frozen finite-tail diagnostic makes 17 `SAFE` predictions across 28 held-out cells with zero observed false-safes, and recovers the same ordering in a subsequent blind million-document evaluation. Finally, we show that deployed ANN error separates into intrinsic source–target candidate error and an additional index-approximation penalty, explaining why an otherwise compatible migration can still fail under low-fidelity ANN search.

Contributions. We (i) formulate *candidate compatibility* using candidate gap, migration depth, and neighborhood containment; (ii) validate its structure across 63 development settings, frozen BRIGHT, and controlled Natural Questions experiments from 100K to one million documents; (iii) introduce a frozen, label-free finite-tail diagnostic that does not require native target retrieval; and (iv) separate intrinsic model compatibility from ANN index fidelity, yielding an operating-region view of embedding migration.

A large body of work on embedding-model upgrades focuses on keeping new representations compatible with old ones. Backward-compatible representation learning constrains a new encoder so that it can still be used with a historical gallery, while later work considers future compatibility, open-set upgrades, stale representations, and partial backfills (Shen et al., 2020; Ramanujan et al., 2022; Zhang et al., 2022; Hu et al., 2022; Jaeckle et al., 2023; Cai et al., 2023). A separate line of work takes a post-hoc approach, mapping between already trained embedding spaces with linear transformations or learned adapters (Conneau et al., 2018; Maystre et al., 2025; Yang et al., 2025; Vejendla, 2025; Maekawa et al., 2026).

Candidate compatibility is complementary to both of these directions. We do not retrain either model, and we do not transform one embedding space into the other. Instead, the legacy space is used only to propose documents, after which the target model scores those documents in its own native space. The question is therefore not whether the two spaces can be made interoperable, but whether the old retrieval function is already useful enough to support the new model.

The mechanism we use is, of course, closely related to standard multi-stage retrieval (Karpukhin et al., 2020; Nogueira & Cho, 2019; Khattab & Zaharia, 2020), where a first-stage retriever produces candidates for a more expensive downstream model. Candidate depth is also known to trade recall against downstream computation (Macdonald & Tonellotto, 2021). The distinction in our setting is what the first-stage retriever represents. In a conventional cascade, the stages are chosen as parts of the same retrieval system. Here, the first stage is a historical index that already exists, while the second stage is a future model chosen independently and potentially much later. Our goal is to characterize when those two frozen retrieval functions remain compatible for migration, how that compatibility changes with scale, and how ANN error changes the deployment regime.

Table 1 makes the distinction explicit. Candidate compatibility uses the familiar retrieve–then–rerank primitive, but the object being studied is different: whether a historical retriever that was never trained, aligned, or selected for the new target already exposes a sufficient candidate set. The

Table 1: **How candidate-compatible migration differs from adjacent approaches.** The reranking primitive itself is standard; the distinction is the migration question being studied.

Approach	Primary goal and distinction
Backward-compatible training	Train a new encoder so that its representations remain compatible with a legacy gallery. Compatibility is introduced *during training* rather than measured post hoc.
Post-hoc alignment / adapters	Learn a transformation between independently trained embedding spaces so that vectors from one space can be compared directly with vectors from another.
Partial backfill / FastFill	Reduce or defer the cost of replacing an index by selectively updating, correcting, or materializing new representations.
Multi-stage retrieval	Use a first-stage retriever to generate candidates for a more expensive second-stage scorer, primarily to improve ranking efficiency.
Candidate compatibility	**Ask whether an already deployed, frozen legacy retriever is itself a sufficient candidate substrate for a newly selected target model.** No cross-space transformation is required: the source controls exposure and the target scores candidates natively. We measure this property through candidate gap, migration depth, finite-tail behavior, and ANN fidelity.

contribution is therefore not a new cascade architecture. It is a framework for *measuring, diagnosing, and validating* this post-hoc migration property.

2 Candidate Compatibility for Embedding Migration

Let $\mathcal{D}$ be a corpus indexed by a source encoder S, and let T be the target encoder we want to deploy. For a query q, the source model first retrieves a candidate set

$$C_K^S(q) = \mathrm{TopK}_{d\in\mathcal{D}}\, s_S(q,d), \tag{1}$$

and the target model then reranks only those candidates using its own similarity function $s_T(q,d)$.

Let M_T denote the quality of native target retrieval over the full corpus, and let $M_{T|S_K}$ denote the quality obtained when the target is restricted to the source top-K. We define their difference as the *candidate gap*:

$$\boxed{G(K) = M_T - M_{T|S_K}.} \tag{2}$$

A smaller gap means that the source candidate set preserves more of the target's retrieval quality. In some cases $G(K)$ can even be negative, since restricting retrieval to the source candidates can act as a relevance filter under finite relevance judgments. Throughout the paper, we use nDCG@10 and treat $G(K)$ as a continuous effect size rather than a binary notion of compatibility. We use $\epsilon = 0.01$ as a deliberately strict reference margin, not as a universal deployment threshold.

To summarize how deep the source retrieval must go, we define the *migration depth*

$$K_\epsilon^*(S\to T) = \min\{K : G(K) \le \epsilon\}. \tag{3}$$

This is the smallest source candidate depth at which target reranking reaches the chosen near-target margin. We also measure target-neighborhood containment,

$$\frac{|C_k^T(q) \cap C_K^S(q)|}{k},$$

which captures a different aspect of compatibility: whether the source exposes the documents that appear in the target neighborhood, rather than whether it preserves relevance quality. Full definitions and the exact recovery condition are given in Appendix B.

Finite-tail diagnostic. There is an immediate practical difficulty with measuring K_ϵ^* directly: doing so appears to require native target retrieval, which is precisely what we would like to avoid before a target index exists. T2-v1 instead looks at how the target reranking changes as the source candidate pool is expanded. For a fixed finite pool $K_{\max} = 500$, its core residual-tail statistic is

$$\mathrm{RTR}_{K;K_{\max}}(q) = 1 - \frac{|R_{T|S_K}^{(10)}(q) \cap R_{T|S_{K_{\max}}}^{(10)}(q)|}{10}. \tag{4}$$

Intuitively, this measures how much of the target top-10 at depth K is still changing when the source pool is expanded to $K_{\max}$. The frozen T2-v1 rule combines this quantity with predefined late-arrival statistics and outputs `SAFE`, `EXPAND`, or `UNSAFE_OR_UNCERTAIN`. Importantly, it uses no qrels, candidate-gap labels, or native target rankings. The full frozen rule and its evaluation chronology are provided in Appendix E.

Model compatibility versus index fidelity. So far, $G(K)$ assumes exact source retrieval. In practice, however, the legacy index will often use approximate nearest-neighbor search. Let ANN configuration A return approximate source candidates $\widetilde{C}_K^{S,A}$, and let $M_{T|\widetilde{S}_{K,A}}$ denote the resulting target-reranked quality. We define the deployed candidate gap as

$$G_A(K) = M_T - M_{T|\widetilde{S}_{K,A}}. \tag{5}$$

Adding and subtracting the exact-source candidate quality gives

$$\boxed{G_A(K) = \underbrace{G(K)}_{\text{model-transition gap}} + \underbrace{P_A(K)}_{\text{index approximation penalty}},} \tag{6}$$

where

$$P_A(K) = M_{T|S_K} - M_{T|\widetilde{S}_{K,A}}.$$

This decomposition separates two different sources of error. The first comes from the source–target migration itself: even an exact source index may fail to expose enough useful target candidates. The second comes from ANN approximation, which can degrade an otherwise compatible source–target pair.

For a quality tolerance ϵ, these two terms define the deployment operating region

$$\mathcal{R}_\epsilon(S,T) = \{(K,A) : G(K) + P_A(K) \le \epsilon\}. \tag{7}$$

This makes it possible to distinguish an intrinsically candidate-limited migration from a compatible model pair that is simply being served with an insufficiently accurate ANN configuration.

3 Experimental Setup

Development sweep and frozen BRIGHT evaluation. We begin with a broad development sweep covering nine source–target model pairs across SciFact, NFCorpus, ArguAna, NanoArguAna, NanoFiQA, NanoMSMARCO, and NanoQuora, giving 63 migration settings in total. The source and target models span the MiniLM, Qwen3, GTE, E5, MPNet, and Nomic families. We evaluate candidate depths $K \in \{10, 20, 50, 100, 200, 500\}$ where available. The complete model matrix, checkpoints, prompting details, and per-setting results are reported in the appendix.

After development, we freeze the evaluation protocol and move to BRIGHT (Su et al., 2024). This evaluation contains 522 queries over 413,234 documents. We fix Qwen3-Embedding-8B (Zhang et al., 2025) as the target model and use MiniLM (Wang et al., 2020), Qwen3-0.6B, and Qwen3-4B as source retrievers.

Scaling to one million documents. To test whether the same behavior survives at larger corpus sizes, we build deterministic nested Natural Questions collections from BEIR (Thakur et al., 2021):

$$100\text{K} \subset 250\text{K} \subset 500\text{K} \subset 1\text{M}.$$

We use 500 unlabeled queries for the T2-v1 probe and a separate set of 1,000 queries for retrieval evaluation. Exact source and target retrieval are evaluated at $K \in \{20, 50, 100, 200, 500, 1000\}$. The T2-v1 predictions are written and hashed before the native target results are added to the evaluation.

Alignment baseline and ANN evaluation. At the one-million-document scale, we also compare candidate-compatible migration against a simple post-hoc alignment baseline. For each source model, we fit a ridge map from target representations into the corresponding legacy source space using up to 100K paired documents, without relevance labels.

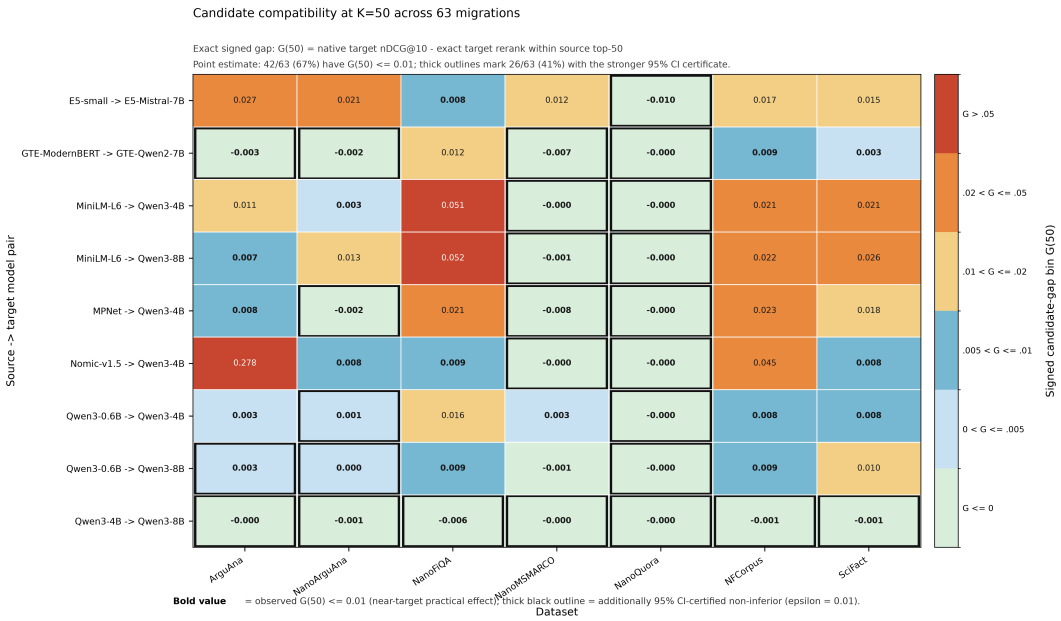

Figure 1: **Candidate compatibility across 63 source–target migrations.** Each cell reports the candidate gap $G(50) = M_T - M_{T|S_{50}}$ in nDCG@10. Smaller values indicate stronger compatibility. Bold cells satisfy the strict reference criterion $G(50) \leq 0.01$ by point estimate; outlined cells additionally satisfy the paired-bootstrap non-inferiority certificate. Compatibility is common but strongly transition-dependent, spanning both near-lossless and substantially candidate-limited migrations.

We then study what happens when the legacy source index is approximate rather than exact. For this experiment, we build FAISS IVF-Flat indexes with $\mathrm{nprobe} \in \{4, 16, 64, 128, 256, 512\}$ and measure source Recall@500 against exact source retrieval before applying target reranking. We report paired-bootstrap non-inferiority intervals with a 0.01 margin. Additional statistical details, the Benjamini–Hochberg audit, and the exact encoding protocols are given in the appendix.

Evaluation chronology and leakage control. The 63-cell sweep is used for development, after which the BRIGHT protocol and T2-v1 thresholds are frozen. T2-v1 is evaluated on 28 held-out migration cells without retuning. In the million-document NQ study, the 500 probe and 1,000 evaluation queries are disjoint, and T2-v1 predictions are persisted and cryptographically hashed before native target retrieval and relevance labels are joined. ANN experiments subsequently vary only source-index fidelity while keeping the model-pair diagnostic fixed.

4 Results

4.1 Broad Evaluation and Frozen Validation

We first ask how often a legacy retriever is already good enough to support a new target model at a modest candidate depth. At $K = 50$, 42 of the 63 development settings (67%) are within 0.01 nDCG@10 of native target retrieval by point estimate. The stricter paired-bootstrap non-inferiority test certifies 26 of the 63 settings. Even after removing NanoQuora and NanoMSMARCO, which are the two most saturated Nano tasks, 25 of the remaining 45 migrations still fall within the same 0.01 margin. Candidate compatibility is therefore fairly common in our sweep, but it is clearly not universal. For that reason, we treat the full $G(K)$ curve as the primary description rather than reducing each migration to a single binary label.

Figure 1 also makes clear that compatibility is not tied to one particular model family or dataset. Some transitions are consistently easy: Qwen3-4B→8B, for example, requires only shallow candi-

Table 2: **Frozen BRIGHT validation at $K = 50$.** The same migration ordering observed during development reappears on a reasoning-intensive external benchmark.

Source → 8B	$M_{T\|S_{50}}$	$G(50)$	**Contain.@10**
MiniLM → 8B	0.1459	0.0367	50.3%
Qwen3-0.6B → 8B	0.1679	0.0146	75.9%
Qwen3-4B → 8B	**0.1790**	**0.0035**	**88.0%**

date sets across the development tasks. Other, especially cross-family, transitions show much larger gaps. This is why we view compatibility as a property of the *source–target transition*, rather than something that can be assigned to the source or target encoder alone.

The overall pattern is not driven only by the easiest Nano tasks. Removing NanoQuora lowers the point-estimate compatibility rate from 66.7% to 61.1%. Removing both NanoQuora and NanoMS-MARCO lowers it further, but 25/45 migrations (55.6%) still remain within the strict 0.01 reference margin. As expected, the confidence-based criterion is more conservative: 11/45 settings are certified in this strictest subset. We therefore describe candidate compatibility as *common but transition-dependent*, rather than as a property that should be expected for every model pair.

We next check whether the ordering found during development survives on a frozen external benchmark. It does. On BRIGHT, native Qwen3-8B obtains 0.1825 nDCG@10, and the three source models again form a clear hierarchy.

At $K = 50$, MiniLM candidates leave a gap of 0.0367, Qwen3-0.6B reduces that gap to 0.0146, and Qwen3-4B is already within 0.0035 of native target retrieval. The resulting point-estimate migration depths satisfy

$$K^*_\epsilon(\text{4B}\to\text{8B}) < K^*_\epsilon(\text{0.6B}\to\text{8B}) < K^*_\epsilon(\text{MiniLM}\to\text{8B}).$$

In other words, the amount of legacy retrieval needed for the target model follows the same ordering on BRIGHT as it did during development. Full per-domain results and confidence intervals are given in the appendix.

4.2 The Migration Hierarchy Persists to One Million Documents

We then test whether this hierarchy survives when the candidate universe becomes much larger. On the one-million-document Natural Questions corpus, the same ordering remains clear. Qwen3-4B→8B is effectively lossless at $K = 50$, Qwen3-0.6B→8B has a small remaining gap and enters the strict reference regime by $K = 200$, while MiniLM→8B requires $K = 500$.

The more interesting result is how little this structure changes as the corpus grows. At $K = 50$, MiniLM has gaps of 0.0278, 0.0327, 0.0315, and 0.0326 as the corpus grows from 100K to 1M documents. Qwen3-0.6B stays in a narrow range around 0.017–0.019, while Qwen3-4B remains very close to zero throughout. Larger corpora make retrieval harder overall, but they do not erase the differences between the three source–target transitions.

The same ordering also survives the stronger paired-bootstrap criterion at one million documents. The CI-certified migration depths are $K = 20$ for Qwen3-4B→8B, $K = 500$ for Qwen3-0.6B→8B, and $K = 1000$ for MiniLM→8B.

The results also highlight an important distinction between neighborhood overlap and retrieval quality. Target top-10 containment inside the source top-50 rises from 70.1% for MiniLM to 95.4% for Qwen3-4B, while the corresponding $G(50)$ values range from 0.0326 to effectively zero. These quantities are related, but they are not the same: containment measures how much of the target neighborhood the source exposes, whereas $G(K)$ measures whether that exposed set is sufficient to preserve relevance quality.

Taken together with the development and BRIGHT results, the scale experiment suggests that migration difficulty is fairly stable for a given source–target transition across the corpus sizes we study. The hierarchy does not disappear when moving from small benchmark collections to nested NQ corpora containing up to one million documents.

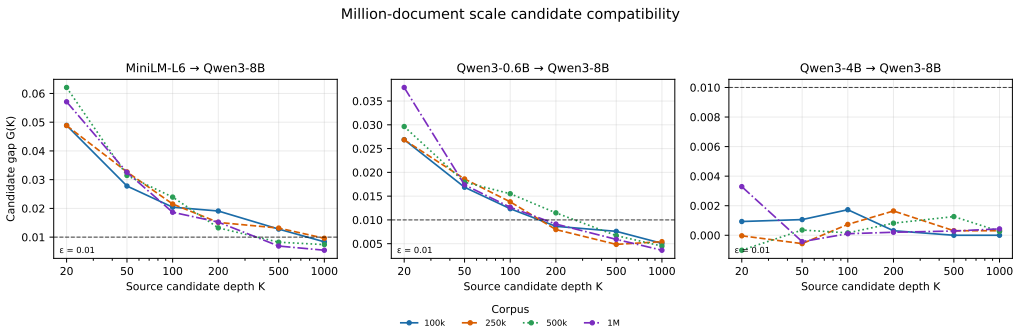

Figure 2: **Candidate compatibility persists as corpus scale grows.** Candidate gap $G(K)$ across source candidate depth on nested Natural Questions corpora from 100K to 1M documents. The migration hierarchy remains stable across scale: Qwen3-4B→8B converges at shallow depth, Qwen3-0.6B→8B is intermediate, and MiniLM→8B requires deeper candidate pools.

Table 3: **Frozen T2-v1 holdout evaluation.** The rule is frozen before these evaluations. "Compatible" denotes $G(50) \leq 0.01$; false-safe means T2-v1 predicts `SAFE` for an incompatible transition.

Migration group	Cells	Compatible	SAFE	False-safe
Qwen3-0.6B → 4B	7	6	4	0
MiniLM → 8B	7	3	2	0
Qwen3-0.6B → 8B	7	6	4	0
Qwen3-4B → 8B	7	7	7	0
Total	**28**	**22**	**17**	**0**

4.3 A Frozen Finite-Tail Probe Predicts Compatibility

We next ask whether compatibility can be estimated before native target retrieval is available. The frozen T2-v1 rule is evaluated on 28 held-out migration cells spanning both source-model and target-model shifts.

Of the 28 held-out cells, 22 satisfy the strict $G(50) \leq 0.01$ criterion. T2-v1 marks 17 of those as `SAFE`, corresponding to 77.3% coverage of the compatible cases. All six incompatible cells are deferred rather than certified. Put differently, every one of the 17 `SAFE` decisions is correct on this frozen holdout set. We describe this as *zero observed false-safes*; it should not be read as a guarantee that future evaluations will also have zero error.

The same frozen rule is then carried into the blind one-million-document NQ experiment. It labels MiniLM→8B and Qwen3-0.6B→8B as `UNSAFE_OR_UNCERTAIN`, while Qwen3-4B→8B is labeled `SAFE`. These decisions agree with the regimes obtained later from exact target retrieval.

Probe-size sensitivity is favorable: 250-query subsamples reproduce the full positive-control decision in 99.5% of trials, while the two harder migrations remain deferred at every tested size. Full results are in Appendix E.

The main point is that finite-pool tail behavior contains useful information about migration compatibility even when native target retrieval has not been constructed. T2-v1 remains an empirical diagnostic rather than a formal certificate; the full rule, frozen holdout ledger, and additional sample-size results are given in Appendix E.

4.4 Ridge Alignment Does Not Explain the Effect

One possible explanation for the results is that the old and new spaces are simply easy to align. We test this with a ridge baseline at one million documents. The best ridge maps obtain 0.5274, 0.6354, and 0.6754 nDCG@10 for MiniLM, Qwen3-0.6B, and Qwen3-4B, respectively. Target reranking of only the source top-50 reaches 0.6486, 0.6638, and 0.6816.

Table 4: **Probe sample-size sensitivity on the 1M NQ probe set.** Entries show agreement with the full-500 decision for the positive Qwen3-4B→8B transition. The two difficult migrations remain deferred in every evaluated subsample.

Probe queries	25	50	100	250	500
4B→8B matches full SAFE	71.0%	74.5%	89.5%	99.5%	100%

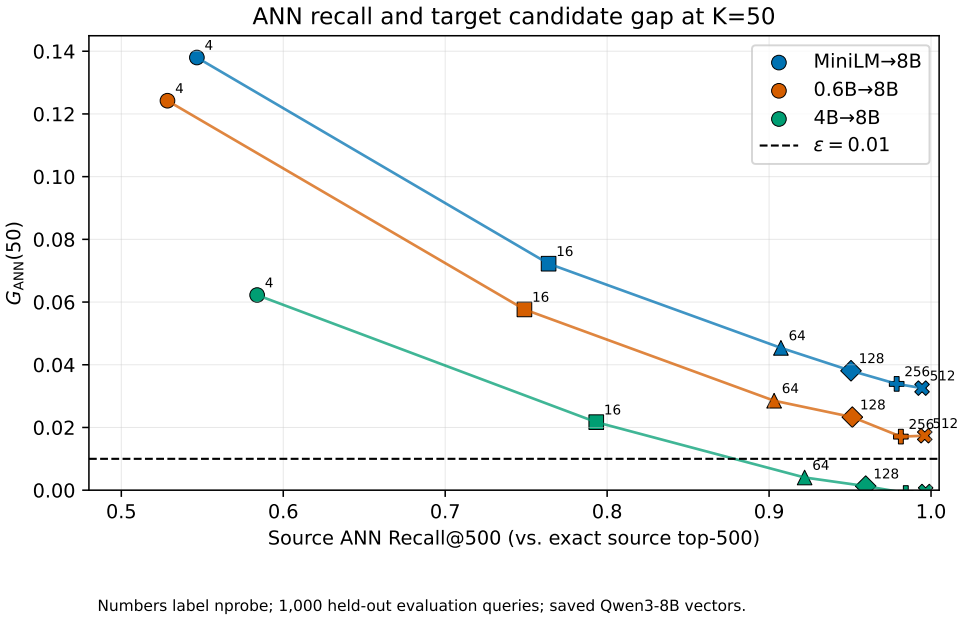

Figure 3: ANN source Recall@500 versus $G_A(50)$ on one million documents. High-recall ANN converges to each transition's exact candidate gap; the Qwen3-4B→8B transition occupies a substantially broader near-target operating region.

Thus, useful retrieval-level compatibility remains even when this simple attempt at direct coordinate interoperability recovers less of the target quality. We intentionally limit this comparison to ridge alignment; these results are not meant to establish superiority over all possible learned alignment or adapter methods.

4.5 ANN Validates the Compatibility–Fidelity Decomposition

The experiments above use exact source retrieval, but a deployed system will usually rely on approximate nearest-neighbor search. We therefore vary the fidelity of the source index and measure how the resulting migration gap changes.

The behavior follows the decomposition $G_A(K) = G(K) + P_A(K)$ closely. As ANN recall increases, each migration approaches its exact-source candidate gap. For Qwen3-4B→8B, $\mathrm{nprobe} = 64$ is already enough to reach $G_A(50) = 0.0040$ at 92.19% source Recall@500. In other words, this transition remains within the strict near-target criterion despite losing roughly 8% of the exact source top-500.

MiniLM and Qwen3-0.6B behave differently. Even when ANN recall approaches 100%, both remain outside the $K = 50$ margin because their exact-source $G(50)$ values are already larger than 0.01. Increasing ANN quality cannot remove an intrinsic candidate gap; these migrations instead need deeper candidate pools.

The required ANN fidelity also differs substantially across migrations. Among the IVF configurations we test, the first setting to reach the strict criterion uses 92.19% source Recall@500 for Qwen3-4B→8B at $K = 20$, 98.14% for Qwen3-0.6B→8B at $K = 200$, and 97.88% for MiniLM→8B at $K = 500$. These values are observed operating points for our tested configurations, not universal recall thresholds.

This experiment also clarifies the low-recall T2-v1 false-safes observed under ANN retrieval. T2-v1 measures the intrinsic model-transition term $G(K)$; it does not measure the additional index penalty $P_A(K)$. A deployment decision therefore needs to check two separate things: whether the model pair is candidate-compatible and whether the source ANN index is accurate enough to preserve that compatibility. Full IVF and HNSW results are reported in the appendix.

5 Deployment Implications and Scope

The operating-region view gives three practical knobs: candidate depth K, ANN search effort, and target-vector materialization. Easy migrations such as Qwen3-4B→8B work with shallow candidate sets and tolerate moderate ANN error; harder transitions require deeper retrieval, higher ANN fidelity, or more backfilling.

Candidate compatibility also allows target vectors to be materialized as they are encountered rather than for the entire corpus up front. This turns migration into a workload-driven process, although it does not guarantee lower lifetime encoding cost. We give a simple cold-start cost model, cache simulations, and target-budget experiments in the appendix.

Our large-scale study fixes Qwen3-8B as the target and varies three source contracts through one million documents, while the 63-cell sweep provides broader model-family coverage. The resulting hierarchy should therefore be read as a property of the audited source–target contracts, not a causal law of parameter count. Full encoding contracts and sensitivity analyses are given in the appendix.

We also measure serving latency in a separate 25K-document H100 benchmark. Native Qwen3-8B retrieval has p50/p95 latency of 18.98/21.56 ms, while warm Qwen3-4B→8B retrieval at $K = 50$ and $\mathrm{nprobe} = 64$ takes 38.79/42.69 ms. Of the 19.81 ms added median latency, 19.49 ms comes from the legacy query encoder; ANN search, cache lookup, target scoring, and top-k selection together contribute less than 1 ms. Cold-cache materialization is substantially more expensive, motivating asynchronous encoding or prewarming. Full latency measurements are reported in Appendix L.

6 Conclusion

Embedding-space incompatibility does not necessarily imply retrieval incompatibility. Across 63 development settings, frozen BRIGHT validation, and Natural Questions collections up to one million documents, legacy retrievers often expose enough candidates to recover near-native target quality. The required candidate depth depends strongly on the source–target transition and remains stable across the corpus scales we study.

A frozen finite-tail diagnostic estimates this regime without constructing a native target index, while

$$G_A(K) = G(K) + P_A(K)$$

separates intrinsic candidate compatibility from ANN approximation error. The contribution is not a new reranking architecture, but a framework for deciding when an existing retrieval function can remain useful during an embedding-model upgrade. In favorable cases, this allows the new model to be introduced before a full corpus backfill is complete.

Reproducibility Statement

The appendix provides complete model and encoding contracts, dataset construction details, statistical audits, the frozen T2-v1 specification, scale sweeps, alignment analyses, ANN experiments, latency measurements, cache simulations, and additional sensitivity analyses. An anonymized supplementary archive provides the frozen experimental specification, machine-readable summaries of

the main results, and lightweight verification scripts that reproduce the principal reported quantities without requiring a GPU or network access.

AI USE STATEMENT

Generative AI tools were used for code generation, debugging, parts of the experimental workflow, and language editing. All claims, results, analyses, and AI-assisted text were reviewed and verified by the authors.

REFERENCES

Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 183–192, 2023. doi: 10.1145/3583780.3614947.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018.

Weihua Hu, Rajas Bansal, Kaidi Cao, Nikhil Rao, Karthik Subbian, and Jure Leskovec. Learning backward compatible embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3018–3028, 2022. doi: 10.1145/3534678.3539194.

Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. In *The Eleventh International Conference on Learning Representations*, 2023.

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*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 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.

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 *NeurIPS*, 2020.

Craig Macdonald and Nicola Tonellotto. On approximate nearest neighbour selection for multi-stage dense retrieval. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 3318–3322, 2021. doi: 10.1145/3459637.3482156.

Seiji Maekawa, Moin Aminnaseri, Pouya Pezeshkpour, and Estevam Hruschka. Align then train: Efficient retrieval adapter learning. *arXiv preprint arXiv:2604.03403*, 2026.

Lucas Maystre, Alvaro Ortega Gonzalez, Charles Park, Rares Dolga, Tudor Berariu, Yu Zhao, and Kamil Ciosek. When embedding models meet: Procrustes bounds and applications. *arXiv preprint arXiv:2510.13406*, 2025.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19386–19395, 2022.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *CVPR*, 2020. doi: 10.1109/CVPR42600.2020.00643.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15938–15949. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.805.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, 2020.

Beining Yang, Yang Cao, and Yang Ren. Integrating vector databases across embedding models. *Proceedings of the ACM on Management of Data*, 3(6):1–28, 2025. doi: 10.1145/3769803.

Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Fanzi Wu, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. Towards universal backward-compatible representation learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 1615–1621, 2022. doi: 10.24963/ijcai.2022/225.

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.

A Datasets, Models, and Reproducibility Details

A.1 Development datasets

Our development sweep covers nine source–target migrations across seven retrieval datasets. Table 5 gives the number of evaluation queries and documents used in the stored results. SciFact, NFCorpus, and ArguAna come from BEIR (Thakur et al., 2021); the Nano tasks are the fixed compact retrieval variants used throughout development.

Table 5: Development dataset statistics.

Dataset	Evaluation queries	Documents
SciFact	300	5,183
NFCorpus	323	3,633
ArguAna	238	8,674
NanoArguAna	50	3,635
NanoFiQA	50	4,571
NanoMSMARCO	50	5,043
NanoQuora	50	5,046

A.2 Frozen BRIGHT evaluation

For the frozen BRIGHT evaluation (Su et al., 2024), we use 522 queries drawn from four domains. Qwen3-Embedding-8B (Zhang et al., 2025) is fixed as the target model. Each source model retrieves its candidates independently, after which those candidates are reranked using exact similarity in the target space.

Table 6: Frozen BRIGHT evaluation statistics.

Domain	Evaluation queries	Documents	Qrel rows
Biology	103	57,359	372
Stack Overflow	117	107,081	478
TheoremQA	194	188,002	617
Sustainable Living	108	60,792	576

A.3 Million-document Natural Questions construction

For the large-scale experiment, we build a deterministic one-million-document subset of Natural Questions from BEIR (Thakur et al., 2021). We select 1,500 queries in total: 500 for the T2-v1 probe and a disjoint set of 1,000 for held-out retrieval evaluation. The evaluation split contains 1,833 qrel rows. The 500 probe queries are used only by the frozen T2-v1 diagnostic and never enter the held-out retrieval-quality calculation.

To study scale while keeping the collections directly comparable, we construct deterministic nested corpora of 100K, 250K, 500K, and 1M documents. Any document required by the selected relevance judgments is retained, and the remaining slots are filled using the deterministic corpus-selection procedure in the experiment package. Thus

$$\mathcal{D}_{100K} \subset \mathcal{D}_{250K} \subset \mathcal{D}_{500K} \subset \mathcal{D}_{1M}.$$

We do not evaluate the full BEIR NQ corpus. All of our large-scale claims therefore refer specifically to this controlled study up to one million documents.

Before native target retrieval was joined to the evaluation, we wrote the blind T2-v1 prediction file to disk and hashed it. Its SHA256 digest is

`d6b9007a5ac2ff42fe346ca56ce5d8b8d52197edf0f395cfef26bdf7c7261470`.

A.4 EMBEDDING MODELS AND STORAGE

Table 7: Embedding model configurations used in the development sweep.

Model	Checkpoint	Pooling	Stored precision	Dimension
MiniLM-L6	`sentence-transformers/all-MiniLM-L6-v2`	Mean	FP16	384
Qwen3-Embedding-0.6B	`Qwen/Qwen3-Embedding-0.6B`	Last token	FP16	1,024
Qwen3-Embedding-4B	`Qwen/Qwen3-Embedding-4B`	Last token	FP16	2,560
Qwen3-Embedding-8B	`Qwen/Qwen3-Embedding-8B`	Last token	FP16	4,096
GTE-ModernBERT	`Alibaba-NLP/gte-modernbert-base`	CLS	FP16	768
GTE-Qwen2-7B	`Alibaba-NLP/gte-Qwen2-7B-instruct`	Last token	FP16	3,584
E5-small-v2	`intfloat/e5-small-v2`	Mean	FP16	384
E5-Mistral-7B	`intfloat/e5-mistral-7b-instruct`	Last token	FP16	4,096
MPNet-base-v2	`sentence-transformers/all-mpnet-base-v2`	Mean	FP16	768
Nomic Embed v1.5	`nomic-ai/nomic-embed-text-v1.5`	Mean	FP16	768

All retrieval vectors are L2-normalized, so the dot products used in the saved large-scale candidate tables are equivalent to cosine similarity. T2-v1 does not calibrate scores across models: the frozen specification uses only saved source ranks and scores together with target scores computed inside the finite source candidate pool.

A.5 AUDITED LARGE-SCALE MODEL CONTRACTS

For the million-document study, we retain the exact checkpoint revisions and encoding contracts for all four models. We reproduce those contracts here because small implementation details—including query prompts, padding, pooling, and truncation—can change retrieval behavior enough to matter in a migration experiment.

Table 8: Audited encoding contracts for the models used in the million-document experiment. Exact revision identifiers are retained for reproducibility.

Model	Revision	Pooling	Padding / truncation	Max length	Query instruction	Normalization
Qwen3-8B	`1d8ad4ca9b3dd8059ad90a75d4983776a23d44af`	Last non-pad	Left / right	8192	Exact template below	FP32 L2
Qwen3-4B	`5cf2132abc99cad020ac570b19d031efec650f2b`	Last non-pad	Left / right	8192	Same Qwen3 contract	FP32 L2
Qwen3-0.6B	`97b0c614be4d77ee51c0cef4e5f07c00f9eb65b3`	Last token	Left / retained contract	512	Historical trailing-space prompt	FP32 L2
MiniLM-L6	`1110a243fdf4706b3f48f1d95db1a4f5529b4d41`	Mean	Model default	Model default	None	L2

For Qwen3-Embedding-8B, the exact query template is

```
Instruct:  Given a web search query, retrieve relevant
passages that answer the query\nQuery:{text}
```

The document instruction is empty. Qwen3-Embedding-4B uses the same verified Qwen3 contract. For Qwen3-Embedding-0.6B, we deliberately preserve the historical source-cache contract, including the trailing space after `Query:  `, rather than changing the source encoding midway through the scale study. MiniLM uses no prefix. All saved retrieval vectors are normalized, so inner product is equivalent to cosine similarity. Vectors are stored in FP16 except for explicit verification samples that request FP32.

The broader 63-cell development sweep uses the checkpoints, pooling strategies, dimensions, and stored precisions listed in Table 7. A machine-readable version of the protocol audit is included with the supplementary materials.

B COMPLETE DEVELOPMENT-SWEEP RESULTS

For these tables, the signed candidate gap is $G(50) = M_T - M_{T|S_{50}}$. The "$G \leq .01$" column is computed from the unrounded point estimate. CI-NI denotes the stricter paired-bootstrap non-inferiority certificate used in the development sweep. One sign convention is worth noting: the stored confidence intervals are for $\Delta_T = M_{T|S_{50}} - M_T$, so their sign is the opposite of G.

Table 9: Development results for MiniLM-L6 → Qwen3-4B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6451	0.7768	0.7562	0.0206	[-0.0363,-0.0069]	No	No
NFCorpus	0.3166	0.4075	0.3868	0.0207	[-0.0349,-0.0078]	No	No
ArguAna	0.3560	0.5139	0.5027	0.0112	[-0.0217,-0.0019]	No	No
NanoArguAna	0.5491	0.6286	0.6256	0.0030	[-0.0250,0.0169]	Yes	No
NanoFiQA	0.4775	0.6935	0.6426	0.0509	[-0.1055,-0.0080]	No	No
NanoMSMARCO	0.5540	0.6799	0.6799	0.0000	[0.0000,0.0000]	Yes	Yes
NanoQuora	0.9368	0.9402	0.9402	0.0000	[0.0000,0.0000]	Yes	Yes

Table 10: Development results for MiniLM-L6 → Qwen3-8B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6451	0.7904	0.7641	0.0263	[-0.0463,-0.0092]	No	No
NFCorpus	0.3166	0.4091	0.3872	0.0218	[-0.0360,-0.0087]	No	No
ArguAna	0.3560	0.5209	0.5138	0.0070	[-0.0177,0.0032]	Yes	No
NanoArguAna	0.5491	0.6706	0.6578	0.0128	[-0.0607,0.0184]	No	No
NanoFiQA	0.4775	0.6985	0.6469	0.0515	[-0.1052,-0.0084]	No	No
NanoMSMARCO	0.5540	0.7152	0.7160	-0.0008	[0.0000,0.0022]	Yes	Yes
NanoQuora	0.9368	0.9512	0.9512	0.0000	[0.0000,0.0000]	Yes	Yes

Table 11: Development results for Qwen3-0.6B → Qwen3-4B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6997	0.7768	0.7692	0.0076	[-0.0179,0.0021]	Yes	No
NFCorpus	0.3594	0.4075	0.3998	0.0077	[-0.0162,-0.0005]	Yes	No
ArguAna	0.4716	0.5139	0.5109	0.0030	[-0.0105,0.0028]	Yes	No
NanoArguAna	0.6452	0.6286	0.6273	0.0013	[-0.0046,0.0009]	Yes	Yes
NanoFiQA	0.5556	0.6935	0.6775	0.0160	[-0.0404,0.0035]	No	No
NanoMSMARCO	0.5853	0.6799	0.6773	0.0026	[-0.0258,0.0181]	Yes	No
NanoQuora	0.9347	0.9402	0.9402	0.0000	[0.0000,0.0000]	Yes	Yes

Table 12: Development results for Qwen3-0.6B → Qwen3-8B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6997	0.7904	0.7801	0.0103	[-0.0202,-0.0028]	No	No
NFCorpus	0.3594	0.4091	0.3996	0.0095	[-0.0177,-0.0026]	Yes	No
ArguAna	0.4716	0.5209	0.5182	0.0027	[-0.0085,0.0011]	Yes	Yes
NanoArguAna	0.6452	0.6706	0.6701	0.0005	[-0.0014,0.0000]	Yes	Yes
NanoFiQA	0.5556	0.6985	0.6896	0.0088	[-0.0361,0.0158]	Yes	No
NanoMSMARCO	0.5853	0.7152	0.7160	-0.0007	[-0.0166,0.0176]	Yes	No
NanoQuora	0.9347	0.9512	0.9512	0.0000	[0.0000,0.0000]	Yes	Yes

Table 13: Development results for Qwen3-4B → Qwen3-8B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.7768	0.7904	0.7917	-0.0012	[0.0000,0.0036]	Yes	Yes
NFCorpus	0.4075	0.4091	0.4101	-0.0011	[-0.0004,0.0027]	Yes	Yes
ArguAna	0.5139	0.5209	0.5211	-0.0002	[-0.0015,0.0018]	Yes	Yes
NanoArguAna	0.6286	0.6706	0.6718	-0.0012	[-0.0042,0.0079]	Yes	Yes
NanoFiQA	0.6935	0.6985	0.7042	-0.0058	[0.0000,0.0173]	Yes	Yes
NanoMSMARCO	0.6799	0.7152	0.7152	0.0000	[0.0000,0.0000]	Yes	Yes
NanoQuora	0.9402	0.9512	0.9512	0.0000	[0.0000,0.0000]	Yes	Yes

Table 14: Development results for GTE-ModernBERT → GTE-Qwen2-7B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.7562	0.7916	0.7889	0.0027	[-0.0141,0.0064]	Yes	No
NFCorpus	0.3497	0.4021	0.3930	0.0091	[-0.0214,0.0032]	Yes	No
ArguAna	0.5786	0.4821	0.4847	-0.0026	[-0.0001,0.0061]	Yes	Yes
NanoArguAna	0.7704	0.6962	0.6978	-0.0016	[-0.0038,0.0082]	Yes	Yes
NanoFiQA	0.6198	0.7440	0.7317	0.0123	[-0.0337,0.0042]	No	No
NanoMSMARCO	0.6446	0.7042	0.7114	-0.0071	[0.0000,0.0214]	Yes	Yes
NanoQuora	0.9671	0.9542	0.9542	0.0000	[0.0000,0.0000]	Yes	Yes

Table 15: Development results for E5-small-v2 → E5-Mistral-7B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6807	0.7707	0.7559	0.0148	[-0.0304,-0.0015]	No	No
NFCorpus	0.3196	0.4007	0.3842	0.0165	[-0.0280,-0.0054]	No	No
ArguAna	0.3472	0.4464	0.4196	0.0268	[-0.0451,-0.0089]	No	No
NanoArguAna	0.4766	0.4829	0.4622	0.0207	[-0.0558,0.0118]	No	No
NanoFiQA	0.4303	0.5955	0.5877	0.0078	[-0.0537,0.0384]	Yes	No
NanoMSMARCO	0.6007	0.6658	0.6534	0.0124	[-0.0379,0.0007]	No	No
NanoQuora	0.9048	0.9397	0.9494	-0.0097	[-0.0017,0.0306]	Yes	Yes

Table 16: Development results for MPNet-base-v2 → Qwen3-4B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.6557	0.7768	0.7590	0.0178	[-0.0371,-0.0012]	No	No
NFCorpus	0.3341	0.4075	0.3847	0.0228	[-0.0366,-0.0103]	No	No
ArguAna	0.3689	0.5139	0.5058	0.0081	[-0.0210,0.0032]	Yes	No
NanoArguAna	0.5297	0.6286	0.6306	-0.0020	[0.0000,0.0049]	Yes	Yes
NanoFiQA	0.5840	0.6935	0.6722	0.0213	[-0.0713,0.0114]	No	No
NanoMSMARCO	0.6345	0.6799	0.6882	-0.0083	[0.0000,0.0221]	Yes	Yes
NanoQuora	0.9214	0.9402	0.9402	0.0000	[0.0000,0.0000]	Yes	Yes

Table 17: Development results for Nomic Embed v1.5 → Qwen3-4B at $K = 50$.

Dataset	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI Δ_T	$G \leq .01$	CI-NI
SciFact	0.7019	0.7768	0.7689	0.0078	[-0.0167,-0.0006]	Yes	No
NFCorpus	0.2826	0.4075	0.3625	0.0450	[-0.0616,-0.0296]	No	No
ArguAna	0.0661	0.5139	0.2356	0.2783	[-0.3151,-0.2410]	No	No
NanoArguAna	0.5625	0.6286	0.6206	0.0079	[-0.0248,0.0018]	Yes	No
NanoFiQA	0.4739	0.6935	0.6849	0.0086	[-0.0304,0.0072]	Yes	No
NanoMSMARCO	0.5567	0.6799	0.6799	0.0000	[0.0000,0.0000]	Yes	Yes
NanoQuora	0.9579	0.9402	0.9402	0.0000	[0.0000,0.0000]	Yes	Yes

B.1 Sensitivity to the reference margin and saturated Nano tasks

The 0.01 margin is a reference operating point, not a natural cutoff in the data. To show how sensitive the headline counts are to this choice, we recompute the development classification over several margins. Table 18 is calculated directly from the 63 stored $G(50)$ values above.

Table 18: Sensitivity of the development sweep to the reference margin. Counts are point-estimate cells satisfying $G(50) \leq \epsilon$. Removing one dataset from nine model pairs leaves 54 cells.

ϵ	All 63	Excl. NanoQuora (54)	Excl. NanoQuora + NanoMSMARCO (45)
0	24	15	8
0.0025	26	17	10
0.0050	31	22	14
0.0075	32	23	15
0.0100	42	33	25
0.0150	48	39	30
0.0200	51	42	33

At $\epsilon = .01$, 42/63 cells (66.7%) meet the point-estimate criterion. The count falls to 33/54 (61.1%) without NanoQuora and to 25/45 (55.6%) without both NanoQuora and NanoMSMARCO. The corresponding CI-certified counts are 26/63 (41.3%), 17/54 (31.5%), and 11/45 (24.4%). Removing the most saturated Nano tasks therefore weakens the aggregate result, as expected, but does not remove it.

B.2 Multiple-testing audit

We keep the repeated-comparison audit separate from the paired-bootstrap non-inferiority analysis. The retained development audit contains 1,890 comparisons in total. Under the archived Benjamini–Hochberg procedure, 511 comparisons against the source baseline and 30 against the target baseline are discoveries at $q \leq 0.05$, giving 541 discoveries overall. At $q \leq 0.10$, the total rises to 594.

Table 19: Benjamini–Hochberg audit of the retained development comparisons. The CI non-inferiority flags reported elsewhere are separate quantities and are not BH-adjusted.

Comparison family / threshold	Discoveries	Total comparisons
Against source baseline, $q \leq 0.05$	511	
Against target baseline, $q \leq 0.05$	30	1,890
All retained comparisons, $q \leq 0.05$	541	
All retained comparisons, $q \leq 0.10$	594	1,890

The paired-bootstrap CI-NI flag answers a different question: whether candidate-restricted target retrieval can be certified as lying within the predefined 0.01 non-inferiority margin. These flags are not Benjamini–Hochberg-adjusted discoveries, so we report the two analyses separately throughout the paper.

C Full Frozen BRIGHT Migration Analysis

After development, we froze the BRIGHT protocol without further adjustment. Qwen3-Embedding-8B is the target, and the three source retrievers are MiniLM-L6, Qwen3-Embedding-0.6B, and Qwen3-Embedding-4B.

Table 20: Frozen BRIGHT migration depth for MiniLM-L6 → Qwen3-8B.

K	M_S	M_T	$M_{T\|S_K}$	$G(K)$	95% CI G	$G \leq .01$	CI-NI	$C_{10,K}$
10	0.1132	0.1825	0.1239	0.0586	[0.0406,0.0768]	No	No	0.2559
20	0.1132	0.1825	0.1372	0.0453	[0.0297,0.0617]	No	No	0.3569
50	0.1132	0.1825	0.1459	0.0367	[0.0227,0.0514]	No	No	0.5033
100	0.1132	0.1825	0.1583	0.0242	[0.0120,0.0377]	No	No	0.5946
200	0.1132	0.1825	0.1689	0.0136	[0.0030,0.0254]	No	No	0.6822
500	0.1132	0.1825	0.1707	0.0118	[0.0023,0.0221]	No	No	0.7761

Table 21: Frozen BRIGHT migration depth for Qwen3-0.6B → Qwen3-8B.

K	M_S	M_T	$M_{T\|S_K}$	$G(K)$	95% CI G	$G \leq .01$	CI-NI	$C_{10,K}$
10	0.1318	0.1825	0.1416	0.0410	[0.0279,0.0542]	No	No	0.4429
20	0.1318	0.1825	0.1550	0.0275	[0.0170,0.0383]	No	No	0.5937
50	0.1318	0.1825	0.1679	0.0146	[0.0069,0.0231]	No	No	0.7590
100	0.1318	0.1825	0.1719	0.0106	[0.0038,0.0179]	No	No	0.8418
200	0.1318	0.1825	0.1784	0.0041	[-0.0012,0.0101]	Yes	No	0.9025
500	0.1318	0.1825	0.1795	0.0030	[-0.0009,0.0074]	Yes	Yes	0.9464

Table 22: Frozen BRIGHT migration depth for Qwen3-4B → Qwen3-8B.

K	M_S	M_T	$M_{T\|S_K}$	$G(K)$	95% CI G	$G \leq .01$	CI-NI	$C_{10,K}$
10	0.1596	0.1825	0.1681	0.0144	[0.0058,0.0241]	No	No	0.5575
20	0.1596	0.1825	0.1757	0.0069	[0.0011,0.0130]	Yes	No	0.7326
50	0.1596	0.1825	0.1790	0.0035	[-0.0005,0.0083]	Yes	Yes	0.8797
100	0.1596	0.1825	0.1797	0.0029	[-0.0005,0.0066]	Yes	Yes	0.9356
200	0.1596	0.1825	0.1800	0.0025	[0.0002,0.0054]	Yes	Yes	0.9657
500	0.1596	0.1825	0.1814	0.0012	[-0.0002,0.0034]	Yes	Yes	0.9858

C.1 Per-domain BRIGHT results at $K = 50$

The archive does not contain a full depth sweep for each BRIGHT domain, so we report the retained per-domain results at $K = 50$.

Table 23: Frozen BRIGHT per-domain results at $K = 50$.

Domain	Source → target	M_S	M_T	$M_{T\|S_{50}}$	$G(50)$	95% CI G	CI-NI	$C_{10,50}$
Biology	MiniLM-L6 → 8B	0.1292	0.1617	0.1673	-0.0056	[-0.0212,0.0120]	No	0.6825
Stack Overflow	MiniLM-L6 → 8B	0.1095	0.1709	0.1645	0.0064	[-0.0239,0.0388]	No	0.5085
TheoremQA	MiniLM-L6 → 8B	0.1138	0.2220	0.1314	0.0905	[0.0613,0.1221]	No	0.3912
Sustainable Living	MiniLM-L6 → 8B	0.1010	0.1441	0.1312	0.0129	[-0.0088,0.0352]	No	0.5278
Biology	Qwen3-0.6B → 8B	0.1191	0.1617	0.1553	0.0064	[-0.0075,0.0205]	No	0.8398
Stack Overflow	Qwen3-0.6B → 8B	0.1054	0.1709	0.1571	0.0138	[-0.0023,0.0308]	No	0.7632
TheoremQA	Qwen3-0.6B → 8B	0.1613	0.2220	0.1941	0.0279	[0.0120,0.0458]	No	0.6758
Sustainable Living	Qwen3-0.6B → 8B	0.1194	0.1441	0.1444	-0.0004	[-0.0102,0.0098]	Yes	0.8269
Biology	Qwen3-4B → 8B	0.1424	0.1617	0.1580	0.0037	[-0.0001,0.0113]	No	0.9350
Stack Overflow	Qwen3-4B → 8B	0.1620	0.1709	0.1697	0.0012	[-0.0006,0.0044]	Yes	0.9350
TheoremQA	Qwen3-4B → 8B	0.1826	0.2220	0.2129	0.0091	[-0.0007,0.0202]	No	0.8015
Sustainable Living	Qwen3-4B → 8B	0.1319	0.1441	0.1485	-0.0044	[-0.0112,0.0003]	Yes	0.9074

D Million-Document Natural Questions: Complete Exact-Retrieval Results

D.1 Native source and target quality across corpus size

Table 24: Native source and Qwen3-8B target nDCG@10 across the nested NQ corpora.

Migration	Corpus	Source nDCG@10	Native 8B nDCG@10
MiniLM → 8B	100K	0.74585	0.87999
MiniLM → 8B	250K	0.66299	0.81440
MiniLM → 8B	500K	0.59561	0.75406
MiniLM → 8B	1M	0.52313	0.68117
Qwen3-0.6B → 8B	100K	0.78893	0.87999
Qwen3-0.6B → 8B	250K	0.71801	0.81440
Qwen3-0.6B → 8B	500K	0.65777	0.75406
Qwen3-0.6B → 8B	1M	0.57955	0.68117
Qwen3-4B → 8B	100K	0.85761	0.87999
Qwen3-4B → 8B	250K	0.79392	0.81440
Qwen3-4B → 8B	500K	0.73432	0.75406
Qwen3-4B → 8B	1M	0.66138	0.68117

D.2 Paired-bootstrap confidence intervals

The archive contains all 72,000 per-query rows from the scale study: three migrations, four corpus sizes, six candidate depths, and 1,000 evaluation queries. Using these saved values, we recompute 95% paired-bootstrap confidence intervals with 10,000 resamples per cell; no new retrieval or model inference is required. A cell is CI-NI when the lower 95% bound of $\Delta_K = M_{T|S_K} - M_T$ is at least -0.01, or equivalently when the upper 95% bound on $G(K)$ is at most 0.01.

Table 25: Exact-retrieval scale results for MiniLM-L6 → Qwen3-8B. Confidence intervals are 10,000-resample paired-bootstrap intervals recomputed from the archived per-query outputs.

Corpus	K	$M_{T\|S_K}$	$G(K)$	95% CI G	$C_{10,K}$	CI-NI
100K	20	0.83110	0.04889	[0.03718, 0.06123]	0.5199	No
100K	50	0.85221	0.02778	[0.01892, 0.03727]	0.6510	No
100K	100	0.85957	0.02042	[0.01299, 0.02874]	0.7275	No
100K	200	0.86090	0.01909	[0.01228, 0.02675]	0.7955	No
100K	500	0.86719	0.01280	[0.00719, 0.01910]	0.8647	No
100K	1000	0.87163	0.00837	[0.00393, 0.01352]	0.9080	No
250K	20	0.76558	0.04882	[0.03672, 0.06134]	0.5347	No
250K	50	0.78171	0.03269	[0.02312, 0.04297]	0.6604	No
250K	100	0.79288	0.02153	[0.01387, 0.02974]	0.7447	No
250K	200	0.79929	0.01512	[0.00846, 0.02229]	0.8037	No
250K	500	0.80125	0.01316	[0.00745, 0.01949]	0.8666	No
250K	1000	0.80479	0.00961	[0.00489, 0.01505]	0.9022	No
500K	20	0.69198	0.06208	[0.04885, 0.07575]	0.5468	No
500K	50	0.72260	0.03146	[0.02224, 0.04114]	0.6833	No
500K	100	0.73012	0.02394	[0.01598, 0.03256]	0.7579	No
500K	200	0.74083	0.01323	[0.00730, 0.01978]	0.8186	No
500K	500	0.74582	0.00824	[0.00310, 0.01392]	0.8760	No
500K	1000	0.74666	0.00740	[0.00283, 0.01241]	0.9072	No
1M	20	0.62408	0.05710	[0.04388, 0.07115]	0.5625	No
1M	50	0.64862	0.03255	[0.02243, 0.04296]	0.7007	No
1M	100	0.66251	0.01866	[0.01135, 0.02642]	0.7766	No
1M	200	0.66594	0.01523	[0.00870, 0.02254]	0.8357	No
1M	500	0.67428	0.00690	[0.00255, 0.01169]	0.8937	No
1M	1000	0.67573	0.00544	[0.00151, 0.00991]	0.9227	Yes

Table 26: Exact-retrieval scale results for Qwen3-0.6B → Qwen3-8B. Confidence intervals are 10,000-resample paired-bootstrap intervals recomputed from the archived per-query outputs.

Corpus	K	$M_{T\|S_K}$	$G(K)$	95% CI G	$C_{10,K}$	CI-NI
100K	20	0.85311	0.02688	[0.01831, 0.03612]	0.6406	No
100K	50	0.86312	0.01687	[0.01025, 0.02400]	0.7765	No
100K	100	0.86764	0.01235	[0.00702, 0.01832]	0.8506	No
100K	200	0.87134	0.00865	[0.00452, 0.01347]	0.9020	No
100K	500	0.87239	0.00760	[0.00361, 0.01228]	0.9441	No
100K	1000	0.87492	0.00507	[0.00186, 0.00897]	0.9701	Yes
250K	20	0.78756	0.02684	[0.01848, 0.03582]	0.6535	No
250K	50	0.79579	0.01861	[0.01172, 0.02592]	0.7844	No
250K	100	0.80060	0.01380	[0.00788, 0.02043]	0.8500	No
250K	200	0.80643	0.00797	[0.00367, 0.01293]	0.8989	No
250K	500	0.80955	0.00485	[0.00155, 0.00867]	0.9402	Yes
250K	1000	0.80898	0.00543	[0.00236, 0.00904]	0.9621	Yes
500K	20	0.72442	0.02964	[0.02041, 0.03894]	0.6692	No
500K	50	0.73613	0.01793	[0.01114, 0.02534]	0.7969	No
500K	100	0.73855	0.01551	[0.00964, 0.02208]	0.8593	No
500K	200	0.74257	0.01148	[0.00656, 0.01695]	0.9022	No
500K	500	0.74736	0.00670	[0.00306, 0.01097]	0.9426	No
500K	1000	0.74956	0.00450	[0.00168, 0.00788]	0.9620	Yes
1M	20	0.64332	0.03785	[0.02809, 0.04808]	0.6878	No
1M	50	0.66381	0.01737	[0.01061, 0.02504]	0.8154	No
1M	100	0.66851	0.01266	[0.00735, 0.01863]	0.8709	No
1M	200	0.67205	0.00913	[0.00485, 0.01394]	0.9114	No
1M	500	0.67526	0.00591	[0.00260, 0.00983]	0.9489	Yes
1M	1000	0.67755	0.00362	[0.00106, 0.00681]	0.9644	Yes

Table 27: Exact-retrieval scale results for Qwen3-4B → Qwen3-8B. Confidence intervals are 10,000-resample paired-bootstrap intervals recomputed from the archived per-query outputs.

Corpus	K	$M_{T\|S_K}$	$G(K)$	95% CI G	$C_{10,K}$	CI-NI
100K	20	0.87906	0.00093	[-0.00167, 0.00396]	0.8366	Yes
100K	50	0.87893	0.00106	[-0.00099, 0.00375]	0.9407	Yes
100K	100	0.87826	0.00173	[0.00015, 0.00421]	0.9758	Yes
100K	200	0.87969	0.00031	[-0.00000, 0.00092]	0.9886	Yes
100K	500	0.87999	0.00000	[-0.00000, -0.00000]	0.9962	Yes
100K	1000	0.87999	0.00000	[-0.00000, -0.00000]	0.9985	Yes
250K	20	0.81444	-0.00004	[-0.00358, 0.00391]	0.8467	Yes
250K	50	0.81497	-0.00057	[-0.00301, 0.00236]	0.9446	Yes
250K	100	0.81367	0.00073	[-0.00125, 0.00333]	0.9718	Yes
250K	200	0.81276	0.00164	[0.00014, 0.00419]	0.9861	Yes
250K	500	0.81410	0.00031	[-0.00000, 0.00092]	0.9951	Yes
250K	1000	0.81410	0.00031	[-0.00000, 0.00092]	0.9978	Yes
500K	20	0.75507	-0.00101	[-0.00510, 0.00325]	0.8572	Yes
500K	50	0.75370	0.00036	[-0.00251, 0.00372]	0.9460	Yes
500K	100	0.75390	0.00016	[-0.00177, 0.00277]	0.9730	Yes
500K	200	0.75324	0.00082	[-0.00075, 0.00325]	0.9867	Yes
500K	500	0.75279	0.00126	[-0.00000, 0.00379]	0.9955	Yes
500K	1000	0.75379	0.00026	[-0.00000, 0.00079]	0.9976	Yes
1M	20	0.67788	0.00329	[-0.00044, 0.00735]	0.8692	Yes
1M	50	0.68160	-0.00043	[-0.00224, 0.00155]	0.9539	Yes
1M	100	0.68107	0.00010	[-0.00163, 0.00201]	0.9755	Yes
1M	200	0.68097	0.00020	[-0.00047, 0.00122]	0.9880	Yes
1M	500	0.68089	0.00029	[-0.00039, 0.00129]	0.9951	Yes
1M	1000	0.68074	0.00043	[-0.00000, 0.00129]	0.9979	Yes

D.3 Observed and CI-certified migration depths

Table 28: Point-estimate and confidence-certified migration depths across NQ corpus sizes. ">1000" means no tested depth up to 1000 satisfied the criterion.

Corpus	MiniLM obs.	MiniLM cert.	0.6B obs.	0.6B cert.	4B obs.	4B cert.
100K	1000	>1000	200	1000	20	20
250K	1000	>1000	200	500	20	20
500K	500	>1000	500	1000	20	20
1M	500	1000	200	500	20	20

At one million documents, the three migrations separate clearly. MiniLM first enters the 0.01 point-estimate regime at $K = 500$ and is CI-certified at $K = 1000$. Qwen3-0.6B reaches the point-estimate criterion at $K = 200$ and the CI criterion at $K = 500$. Qwen3-4B satisfies both at the shallowest tested depth, $K = 20$.

D.4 SCALE FIGURES

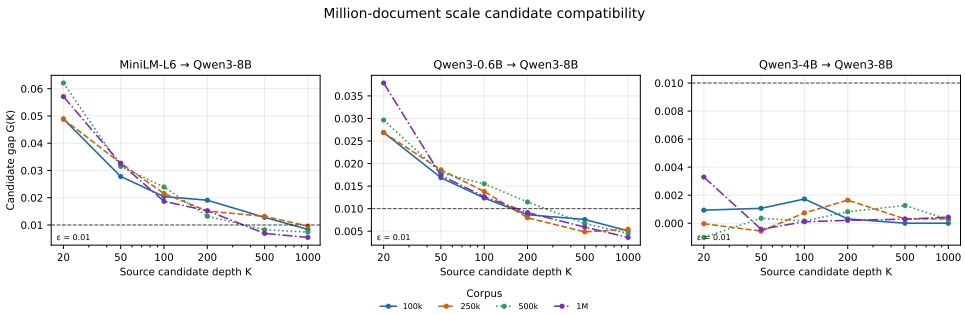

Figure 4: Candidate gap $G(K)$ across all tested depths and nested NQ corpus sizes. Despite the historical filename, this figure plots the full K curve rather than only $K = 50$.

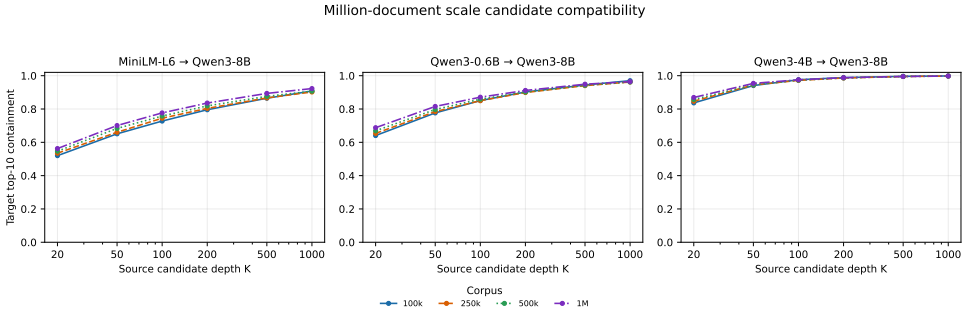

Figure 5: Native Qwen3-8B top-10 containment as a function of source candidate depth and corpus size.

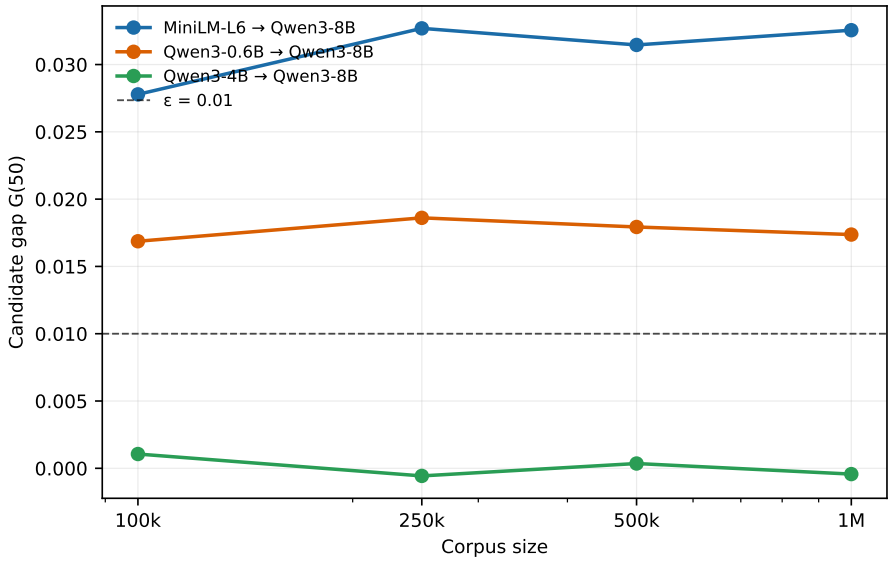

Figure 6: Candidate gap at $K = 50$ as the corpus grows from 100K to one million documents.

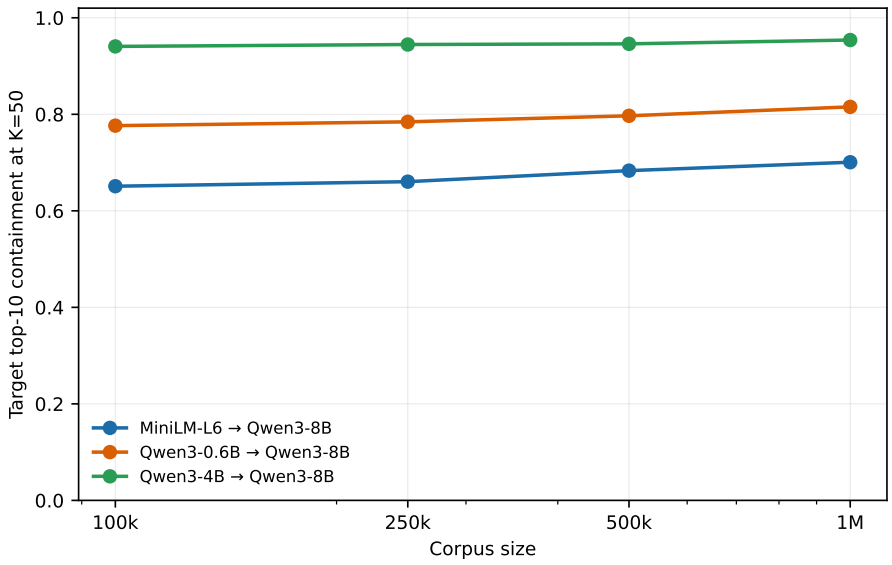

Figure 7: Target top-10 containment inside the source top-50 as corpus size increases.

E Frozen T2-v1 Residual Candidate-Tail Probe

E.1 Freezing chronology and permitted information

T2-v1 was frozen on August 23, 2026, before we inspected the Qwen3-8B target-shift results. The probe is allowed to use source-ranked candidate IDs, source ranks and scores, exact target scores for documents inside the finite source pool, and target rerankings of source prefixes. It is not allowed to use qrels, native full-corpus target rankings, target nDCG, candidate gaps, confidence intervals, or compatibility labels assigned after the fact.

The probe uses R_{500} as its finite reference, meaning target reranking restricted to the source top-500 candidates. This is deliberately not native target retrieval. T2-v1 makes its decision at $K = 50$ using the frozen depth set $K \in \{10, 20, 50, 100, 200, 500\}$.

E.2 Frozen feature definitions

For each query and source prefix:

$$\text{ResidualTail}(K) = 1 - \frac{|\text{Top10}(R_K) \cap \text{Top10}(R_{500})|}{10}, \tag{8}$$

$$\text{Stability}_{500}(K) = \frac{|\text{Top10}(R_K) \cap \text{Top10}(R_{500})|}{10}. \tag{9}$$

Beyond residual tail and stability, the rule uses `deepest_p90`, late-shell entrant probabilities, a weighted late-tail area over the expansions $50 \to 100$, $100 \to 200$, and $200 \to 500$, and two fixed high-risk guards: `last_shell_any_rate` and `fraction_margin_nonpositive`.

E.3 Exact frozen decision rule

At $K = 50$, output `SAFE` iff:

1. mean residual tail ≤ 0.05;
2. `deepest_p90` ≤ 200;
3. late-tail area ≤ 0.20.

If `SAFE` is false, output `UNSAFE_OR_UNCERTAIN` iff at least one guard holds:

- mean residual tail > 0.10;
- mean stability to 500 < 0.90;
- `last_shell_any_rate`> 0.25;
- `fraction_margin_nonpositive`> 0.50.

If neither condition is met, the rule outputs `EXPAND`. Here, `SAFE` has a narrow meaning: it predicts the strict $G(50) \leq 0.01$ criterion. It should not be interpreted as a general production-safety guarantee.

E.4 Blind one-million-document probe results

Table 29: Frozen T2-v1 features and blind decisions on the 500 unlabeled NQ probe queries. Ground-truth $G(50)$ is joined only after the blind file is persisted.

Migration	Residual tail	Stability	Deepest p90	Late-tail area	Last-shell rate	Nonpositive margin	Decision	$G(50)$	False-safe
MiniLM → 8B	0.2452	0.7548	450.2	0.1074	0.462	0.680	UNSAFE_OR_UNCERTAIN	0.03255	No
Qwen3-0.6B → 8B	0.1538	0.8462	399.0	0.0639	0.304	0.490	UNSAFE_OR_UNCERTAIN	0.01737	No
Qwen3-4B → 8B	0.0406	0.9594	137.3	0.0151	0.066	0.196	SAFE	-0.00043	No

E.5 Frozen holdout ledger before the NQ scale study

Before running the one-million-document NQ evaluation, we applied the frozen T2-v1 rule—without retuning—to 28 held-out source–target–dataset cells spanning four migration groups. Table 30 summarizes the resulting ledger.

Table 30: Frozen T2-v1 holdout ledger summarized by migration group. Compatibility uses the strict point-estimate criterion $G(50) \leq 0.01$. "SAFE precision" refers only to the empirically observed frozen holdouts and is not a guarantee.

Holdout group	Cells	Compatible	SAFE	False-safe
Qwen3-0.6B → Qwen3-4B	7	6	4	0
MiniLM → Qwen3-8B	7	3	2	0
Qwen3-0.6B → Qwen3-8B	7	6	4	0
Qwen3-4B → Qwen3-8B	7	7	7	0
Total	28	22	17	0

Among the 22 compatible holdout cells, T2-v1 marks 17 as `SAFE`, giving 77.3% coverage of the compatible cases. All 17 of those `SAFE` decisions are correct on the frozen holdout set, while each of the six incompatible cells is deferred. The observed `SAFE` precision is therefore 17/17, with zero observed false-safes. These numbers describe this finite evaluation only; they are not a guarantee of future precision.

E.6 Probe sample-size sensitivity

We also ask how many unlabeled probe queries are needed for the NQ decision to stabilize. This analysis is exploratory because it repeatedly subsamples the same 500-query NQ probe set rather than introducing a new validation set. T2-v1 remains frozen throughout: no feature, threshold, candidate depth, or decision rule is changed after looking at these trials.

For the positive-control Qwen3-4B→8B migration, agreement with the full 500-query `SAFE` decision increases quickly as more probe queries are used:

Table 31: T2-v1 probe sample-size sensitivity on the fixed NQ probe set. The two more difficult transitions are deferred in all retained smaller-sample trials.

Probe queries	4B→8B matches full-500 decision	MiniLM→8B deferred	0.6B→8B deferred
25	71.0%	100%	100%
50	74.5%	100%	100%
100	89.5%	100%	100%
250	99.5%	100%	100%
500	100%	100%	100%

The effect of smaller probe sets is asymmetric here. Fewer queries mainly make the positive Qwen3-4B→8B decision less stable; MiniLM→8B and Qwen3-0.6B→8B remain deferred in every retained subsample. In this particular experiment, roughly 100–250 unlabeled queries already appear useful for the diagnostic, but we do not generalize that sample-size range beyond the transitions tested here.

F Post-hoc Linear Alignment Baseline

As a simple post-hoc alignment baseline, we fit a ridge map from Qwen3-8B document embeddings into each source document space using paired but unlabeled document representations. At retrieval time, the Qwen3-8B query is mapped into the source space and searched against the unchanged legacy index. The fit uses no relevance judgments.

Table 32: Ridge-alignment train-size sensitivity on the 1M NQ corpus.

Migration	Paired train docs	Mapped nDCG@10	Fit time (s)
MiniLM ← 8B	1,000	0.17598	192.5
MiniLM ← 8B	10,000	0.44104	225.7
MiniLM ← 8B	50,000	**0.52745**	226.5
MiniLM ← 8B	100,000	0.51211	242.2
Qwen3-0.6B ← 8B	1,000	0.36431	234.8
Qwen3-0.6B ← 8B	10,000	0.60532	242.4
Qwen3-0.6B ← 8B	50,000	**0.63538**	252.4
Qwen3-0.6B ← 8B	100,000	0.62800	262.4
Qwen3-4B ← 8B	1,000	0.50227	266.7
Qwen3-4B ← 8B	10,000	0.66528	264.5
Qwen3-4B ← 8B	50,000	**0.67543**	276.8
Qwen3-4B ← 8B	100,000	0.67029	287.0

Table 33: Best ridge alignment versus source-candidate target reranking at one million documents.

Migration	Source	Best ridge	Target rerank, $K = 50$	Native 8B
MiniLM → 8B	0.52313	0.52745	0.64862	0.68117
Qwen3-0.6B → 8B	0.57955	0.63538	0.66381	0.68117
Qwen3-4B → 8B	0.66138	0.67543	0.68161	0.68117

This baseline is intentionally narrow. It tests a simple post-hoc linear alignment and should not be read as evidence that candidate reranking dominates more expressive supervised adapters.

G Approximate Nearest-Neighbor Robustness

G.1 Source-index recall characterization

We measure source Recall@500 against exact source retrieval on the 1,000 held-out NQ evaluation queries.

Table 34: ANN source Recall@500 on the one-million-document corpus.

Source	Index setting	Recall@500
MiniLM	Flat exact	100.00%
MiniLM	HNSW $M = 16$, efSearch=32	39.87%
MiniLM	HNSW $M = 16$, efSearch=64	55.56%
MiniLM	HNSW $M = 16$, efSearch=128	71.51%
MiniLM	HNSW $M = 32$, efSearch=32	45.95%
MiniLM	HNSW $M = 32$, efSearch=64	61.92%
MiniLM	HNSW $M = 32$, efSearch=128	77.06%
MiniLM	IVF-Flat nprobe=4	54.67%
MiniLM	IVF-Flat nprobe=16	76.38%
MiniLM	IVF-Flat nprobe=64	90.73%
Qwen3-0.6B	Flat exact	100.00%
Qwen3-0.6B	HNSW $M = 16$, efSearch=32	37.79%
Qwen3-0.6B	HNSW $M = 16$, efSearch=64	52.90%
Qwen3-0.6B	HNSW $M = 16$, efSearch=128	68.91%
Qwen3-0.6B	HNSW $M = 32$, efSearch=32	43.68%
Qwen3-0.6B	HNSW $M = 32$, efSearch=64	59.07%
Qwen3-0.6B	HNSW $M = 32$, efSearch=128	74.67%
Qwen3-0.6B	IVF-Flat nprobe=4	52.85%
Qwen3-0.6B	IVF-Flat nprobe=16	74.89%
Qwen3-0.6B	IVF-Flat nprobe=64	90.31%
Qwen3-4B	Flat exact	100.00%
Qwen3-4B	HNSW $M = 16$, efSearch=32	40.95%
Qwen3-4B	HNSW $M = 16$, efSearch=64	57.12%
Qwen3-4B	HNSW $M = 16$, efSearch=128	73.89%
Qwen3-4B	HNSW $M = 32$, efSearch=32	46.61%
Qwen3-4B	HNSW $M = 32$, efSearch=64	63.41%
Qwen3-4B	HNSW $M = 32$, efSearch=128	79.13%
Qwen3-4B	IVF-Flat nprobe=4	58.39%
Qwen3-4B	IVF-Flat nprobe=16	79.33%
Qwen3-4B	IVF-Flat nprobe=64	92.19%

For the target-conditioned ANN experiments, we use IVF-Flat with $nlist = 1000$ and nprobe $\in \{4, 16, 64, 128, 256, 512\}$. Target reranking is performed with the saved Qwen3-8B vectors.

Table 35: IVF-Flat ANN-conditioned results for MiniLM-L6 $\rightarrow$ Qwen3-8B on the 1M NQ corpus.

nprobe	Recall@500	$G_A(20)$	$G_A(50)$	$G_A(100)$	$G_A(200)$	$G_A(500)$	$C^A_{10,50}$	K^*_A
4	54.67%	0.15935	0.13800	0.13322	0.13052	0.12729	0.5516	> 500
16	76.38%	0.09920	0.07223	0.06566	0.06251	0.05789	0.6321	> 500
64	90.73%	0.07190	0.04538	0.03546	0.03262	0.02505	0.6766	> 500
128	95.06%	0.06353	0.03808	0.02514	0.02298	0.01454	0.6875	> 500
256	97.88%	0.05845	0.03390	0.01977	0.01659	0.00778	0.6950	500
512	99.45%	0.05710	0.03255	0.01903	0.01523	0.00690	0.6994	500

Table 36: IVF-Flat ANN-conditioned results for Qwen3-0.6B $\rightarrow$ Qwen3-8B on the 1M NQ corpus.

nprobe	Recall@500	$G_A(20)$	$G_A(50)$	$G_A(100)$	$G_A(200)$	$G_A(500)$	$C^A_{10,50}$	K^*_A
4	52.85%	0.13411	0.12422	0.12167	0.12025	0.11984	0.6156	> 500
16	74.89%	0.07318	0.05765	0.05299	0.05244	0.05180	0.7242	> 500
64	90.31%	0.04892	0.02854	0.02455	0.02203	0.02001	0.7848	> 500
128	95.14%	0.04019	0.02331	0.01765	0.01489	0.01181	0.7996	> 500
256	98.14%	0.03500	0.01709	0.01259	0.00936	0.00602	0.8098	200
512	99.61%	0.03685	0.01737	0.01223	0.00913	0.00591	0.8142	200

Table 37: IVF-Flat ANN-conditioned results for Qwen3-4B → Qwen3-8B on the 1M NQ corpus.

nprobe	Recall@500	$G_A(20)$	$G_A(50)$	$G_A(100)$	$G_A(200)$	$G_A(500)$	$C^A_{10,50}$	K^*_A
4	58.39%	0.06488	0.06223	0.06179	0.06199	0.06199	0.7345	> 500
16	79.33%	0.02337	0.02169	0.02158	0.02128	0.02149	0.8596	> 500
64	92.19%	0.00629	0.00405	0.00458	0.00397	0.00434	0.9239	20
128	95.97%	0.00362	0.00129	0.00211	0.00149	0.00157	0.9409	20
256	98.44%	0.00284	-0.00090	-0.00008	-0.00035	-0.00027	0.9491	20
512	99.67%	0.00329	-0.00043	0.00010	0.00020	0.00029	0.9533	20

G.2 Observed ANN operating points

The first tested IVF configuration to reach $G_A(K) \leq 0.01$ looks very different across the three migrations. MiniLM needs 97.878% source Recall@500 and $K^*_A = 500$; Qwen3-0.6B needs 98.137% recall and $K^*_A = 200$; Qwen3-4B reaches the criterion at 92.194% recall with $K^*_A = 20$. These are simply the lowest sufficient recall values among the configurations we tested, not universal ANN thresholds.

At $K = 50$, increasing ANN recall cannot rescue the two harder migrations. MiniLM and Qwen3-0.6B remain outside the strict near-target regime even as recall approaches 100%, with their gaps converging to the corresponding exact-source $K = 50$ values. The remaining error at high recall therefore comes from the source–target candidate relationship itself, not from ANN approximation.

Copy `ann_extended_recall_vs_candidate_gap.pdf` into `paper_appendix_artifacts/`.

Figure 8: ANN source Recall@500 versus target candidate gap. Each migration approaches its exact-retrieval asymptote as ANN recall increases.

G.3 Frozen T2-v1 under ANN candidate pools

Table 38: Frozen T2-v1 predictions on ANN candidate pools for the initial IVF sweep.

Source	nprobe	Recall@500	T2-v1 decision	$G_A(50)$	False-safe
MiniLM	4	54.67%	`UNSAFE_OR_UNCERTAIN`	0.137995	No
MiniLM	16	76.38%	`UNSAFE_OR_UNCERTAIN`	0.072235	No
MiniLM	64	90.73%	`UNSAFE_OR_UNCERTAIN`	0.045385	No
Qwen3-0.6B	4	52.85%	`UNSAFE_OR_UNCERTAIN`	0.124217	No
Qwen3-0.6B	16	74.89%	`UNSAFE_OR_UNCERTAIN`	0.057650	No
Qwen3-0.6B	64	90.31%	`UNSAFE_OR_UNCERTAIN`	0.028536	No
Qwen3-4B	4	58.39%	`SAFE`	0.062234	Yes
Qwen3-4B	16	79.33%	`SAFE`	0.021694	Yes
Qwen3-4B	64	92.19%	`SAFE`	0.004046	No

The two false-safe Qwen3-4B cases make the separation between model compatibility and index fidelity concrete. T2-v1 diagnoses candidate-tail behavior for the model transition; it is not a certificate that the ANN index is healthy.

H Target-Evaluation Budget Analysis

We include this secondary analysis for completeness. Candidate depth K controls which documents are reachable, while the target-evaluation budget B controls how many of those reachable candidates receive exact target scoring. The aggregate follows the retained leakage-safe development procedure and includes 58 cells with a positive, non-negligible all-50 hybrid-gain denominator.

Table 39: Leakage-safe target-evaluation budget curve. Values are medians with IQR across 58 eligible development cells.

B	n	Hybrid med.	Hybrid IQR low	Hybrid IQR high	Native med.	Native IQR
2	58	0.066	-0.014	0.188	0.049	[-0.011,0.183]
4	58	0.227	0.106	0.377	0.212	[0.078,0.359]
8	58	0.434	0.303	0.559	0.395	[0.236,0.595]
12	58	0.585	0.461	0.709	0.551	[0.350,0.715]
16	58	0.682	0.569	0.809	0.631	[0.449,0.800]
20	58	0.775	0.637	0.879	0.693	[0.520,0.877]
24	58	0.818	0.715	0.899	0.744	[0.550,0.895]
32	58	0.902	0.791	0.958	0.801	[0.634,0.989]
40	58	0.937	0.852	0.977	0.816	[0.659,0.945]
50	58	1.000	1.000	1.000	0.932	[0.833,1.051]

With $B = 12$, using 24% of the full 50-candidate target-scoring budget recovers 58.5% of the median all-50 hybrid gain. At $B = 32$, 64% of the budget recovers 90.2%. This is a separate budget analysis; it is not used to define migration depth or T2-v1.

I Progressive-Cache Simulation Details

These experiments simulate workloads; they are not measurements from a production serving system. A cache hit means that the target representation for a document is already resident in the cache, and corpus coverage is counted by documents rather than token mass.

Table 40: Progressive-cache simulation at the earliest available 3M-query horizon for a 10M-document corpus.

Workload	Queries	Target documents	Corpus fraction	Encoding hours
Repeated popularity	3,000,000	7,876,161	0.788	134.29
Uniform	3,000,000	10,000,000	1.000	170.51
Zipf	3,000,000	10,000,000	1.000	170.51

Table 41: Mean cache behavior across the simulated workload traces.

Workload	Trace rows	Mean cache-hit rate	Mean new target encodes/query
Repeated popularity	15	0.954	1.248
Reverse popularity	15	0.550	8.508
Uniform	15	0.550	8.508
Zipf	15	0.770	5.098

Repeated candidate exposure can make a target-vector cache quite effective. Mean hit rates reach 77.0% under Zipf access and 95.4% under repeated-popularity traffic. A high hit rate, however, does not necessarily mean that only a small part of the corpus will ever be encoded. In the retained long-horizon traces, both Zipf and uniform traffic have touched the full 10M-document corpus by the earliest saved 3M-query checkpoint, while repeated-popularity traffic has reached 78.8%. The main benefit of progressive migration is therefore that it avoids requiring a full backfill *up front*; it does not guarantee less total encoding work over the lifetime of the system.

J Analytical Cold-Start Break-Even

The following comparison is analytical rather than a measured throughput result. Let L denote the total corpus token mass, r_S and r_T the source and target document-encoding rates, and L_U the token

Table 43: **Measured warm-path latency on the 25K-document H100 benchmark.**

Configuration	K	p50 (ms)	p95 (ms)
Native Qwen3-8B	–	18.98	21.56
Qwen3-4B→8B	20	38.61	42.38
Qwen3-4B→8B	50	38.79	42.69
Qwen3-0.6B→8B	200	35.32	88.16
MiniLM→8B	500	24.85	90.74

mass of unique documents that are later encoded by the target. A full target backfill costs

$$C_{\text{full}} = \frac{L}{r_T}, \tag{10}$$

whereas a candidate-compatible cold start costs

$$C_{\text{CC}} = \frac{L}{r_S} + \frac{L_U}{r_T}. \tag{11}$$

If we ignore query encoding, ANN construction, batching differences, storage, and hardware effects, the candidate-compatible path uses less cumulative document-encoding time exactly when

$$\boxed{\frac{L_U}{L} < 1 - \frac{r_T}{r_S}.} \tag{12}$$

For example, when $r_S/r_T = 2$, the candidate-compatible path remains below the cost of a full target backfill as long as less than 50% of the corpus token mass has been target-encoded.

K Large-Scale Encoding Throughput and Artifact Provenance

The final large-scale run produced roughly 16GB of stored artifacts. The one-million-document run did not retain reusable FAISS indexes or per-query latency traces, so we do not retrospectively estimate online latency from that experiment. Instead, we measure serving latency separately in a controlled 25K-document H100 benchmark, reported in the following section.

Table 42: Measured encoding-stage wall times for the one-million-document experiment. The Qwen3-8B corpus was encoded in two non-overlapping stages and reused 354,441 probe-candidate vectors.

Model/stage	Documents	Wall time (s)	Approx. docs/s	Notes
MiniLM-L6	1,000,000	1,415.2	706.6	50 shards
Qwen3-0.6B	1,000,000	5,865.8	170.5	50 shards
Qwen3-4B	1,000,000	7,924.6	126.2	50 shards
Qwen3-8B probe pool	354,441	3,348.6	105.8	includes probe-query work
Qwen3-8B remaining corpus	645,559	5,716.1	112.9	reuses probe vectors
Qwen3-8B combined corpus	1,000,000	9,064.7	110.3	no duplicate document encoding

L Serving-Latency Benchmark

We measure online serving latency separately from the one-million-document retrieval-quality experiment. The benchmark uses a 25K-document corpus on an H100 GPU and is intended to characterize the request-path cost of candidate-compatible migration rather than corpus-scale retrieval quality.

Native Qwen3-Embedding-8B retrieval has p50/p95 latency of 18.98/21.56 ms. For the warm Qwen3-4B→8B configuration at $K = 50$ and $\text{nprobe} = 64$, p50/p95 latency is 38.79/42.69 ms, corresponding to an additional 19.81 ms at the median.

Table 44: **Stage-level latency for warm Qwen3-4B→8B at $K = 50$, $\mathrm{nprobe} = 64$.**

Stage	Mean latency (ms)
Source query encoding	19.485
Source ANN search	0.401
Target query encoding	18.743
Candidate + cache lookup	0.213
Target scoring	0.062
Top-k selection	0.067

Table 45: **Cold-cache median latency on the 25K-document H100 benchmark.**

Source	$K = 50$	$K = 200$	$K = 500$
MiniLM-L6	603 ms	2665 ms	5677 ms
Qwen3-0.6B	590 ms	2582 ms	6302 ms
Qwen3-4B	637 ms	2912 ms	6137 ms

For the warm Qwen3-4B→8B, $K = 50$ configuration, the stage-level mean latency is shown in Table 44. The legacy query encoder accounts for almost all of the migration-specific overhead; ANN search, cache lookup, target scoring, and top-k selection together contribute less than 1 ms.

Cold-cache target-document encoding is substantially more expensive. At $K = 50$, median cold-cache latency is approximately 0.59–0.64 s across the three source models; at $K = 200$ it rises to approximately 2.6–2.9 s, and at $K = 500$ to approximately 5.7–6.3 s. These measurements indicate that synchronous materialization of large numbers of missing target vectors is not an attractive request-path strategy. A practical deployment would instead prewarm likely candidates or materialize missing target vectors asynchronously.

At larger candidate depths, p95 latency exhibits intermittent tails that move between individually timed stages under CUDA synchronization. We therefore report the observed p95 values but do not attribute these tails to a specific migration stage.

A simple cache-hit interpolation can be written as

$$L(h, K) \approx L_{\mathrm{warm}}(K) + (1 - h)Kc_{\mathrm{miss}},$$

where h is the cache-hit rate and the measured Qwen3-8B document-encoding penalty is approximately 13.7–15.5 ms per miss. Latencies derived from this expression are modeled estimates rather than direct measurements at each cache hit rate.

M Reproducibility and Leakage Safeguards

- The 500 NQ probe queries and 1,000 held-out evaluation queries are disjoint.
- T2-v1 was frozen before Qwen3-8B target-shift inspection and prohibits qrels, native target retrieval, target nDCG, $G(K)$, confidence intervals, and retrospective labels.
- The exact-retrieval NQ blind predictions were persisted and hashed before native target labels were joined.
- The linear alignment baseline uses paired document embeddings but no relevance labels.
- ANN source fidelity is measured against exact source retrieval; ANN-conditioned target quality is evaluated on the same held-out 1,000 evaluation queries.
- The nested NQ corpus sizes are correlated scale conditions, not independent datasets.
- The 0.01 nDCG@10 margin is a stringent reference operating point; continuous $G(K)$ curves are the primary effect-size description.
- The final frozen holdout ledger contains 28 cells, of which 22 satisfy $G(50) \leq 0.01$; T2-v1 emits 17 `SAFE` decisions with zero observed false-safes in that ledger.

- Probe sample-size sensitivity is labeled exploratory because it resamples the same 500-query NQ probe set rather than introducing a new independent validation set.
- The BH repeated-comparison audit and the paired-bootstrap CI-NI analysis are reported separately; CI-NI flags are not BH-adjusted quantities.
- The one-million-document quality experiment and the 25K-document latency benchmark are separate evaluations. No latency measurements are extrapolated from the one-million-document run.

For reproducibility, the exact frozen T2-v1 specification is provided verbatim in the supplementary material.