A Reliability Audit of Aggregated MTEB Scores: Coverage Asymmetry and Language-Specific Rank Inversions Across 17 Languages

Archit Rastogi
Independent Researcher
architrastogi20@gmail.com

Abstract

Practitioners select multilingual embedding models from aggregated MTEB leaderboard scores, but the same global score can correspond to very different per-language behaviours. We audit MTEB(Multilingual) retrieval rankings against language-specific rankings on **17 languages** across **three nested model rosters** (restricted n=7; deployment-relevant curated n=25; broad MMTEB-eligible n=28–57), using Kendall τ and pairwise inversion rate. At deployment-realistic roster size, median pairwise inversion rate is **33%** (per-language Kendall τ reliable after Benjamini–Hochberg FDR in 11/17 languages); the extended tier shows higher agreement (τ=0.69), reflecting its multilingual-first composition. Per-language task coverage is asymmetric, and inversion rate decreases with the number of MTEB retrieval tasks available for a language (Spearman ρ=−0.59 curated). **A language-balanced aggregator (per-language mean first, then macro-average) drops the median inversion rate from 33% to 14.5%** (leave-one-out, τ=0.71), the lowest among the held-out aggregators we test and our most actionable mitigation. A cluster of *hidden-failure* models, headed by `granite-311m` (Awasthy et al., 2026) (16/17 languages) and `harrier-0.6b` (Microsoft, 2026) (10/17), ranks in the global top half but collapses to the local bottom quartile, with a distinctively shallow English–multilingual NDCG gap. We release the three-tier audit data, a multilingual RAG benchmark (IT/JA/HI × finance/law), and the analysis pipeline.

1 Introduction

The Massive Text Embedding Benchmark (MTEB; Muennighoff et al., 2023), expanded by MMTEB (Enevoldsen et al., 2025), has become the de facto standard for embedding model selection. Although MMTEB exposes language- and task-filtered leaderboard views, the headline aggregated score remains the default user-facing artifact, and practitioners routinely consult it when choosing a model for a specific application. Aggregated scores average performance across hundreds of tasks and dozens of languages, raising the question of how well they proxy language-specific retrieval capability. Prior aggregation critiques have shown that averages across *task types* can hide systematic differences (Colombo et al., 2022; Ethayarajh and Jurafsky, 2020; Dehghani et al., 2021); we extend this analysis to *cross-language* aggregation and show that disagreement between aggregated and language-specific rankings is large enough, at deployment-realistic roster size, to materially affect model selection. Our audit targets the retrieval category of MMTEB(Multilingual), the view most consulted when selecting embedding models for search and RAG; a leaderboard-only audit of classification and STS finds comparable per-language inversions (App. Y), and we scope our own-run validation to retrieval.

Two quantities do the work throughout. **MTEB-agg** is the headline number the leaderboard shows: a model's mean over every retrieval task it was evaluated on. **Lang-avg** is its mean over only the retrieval tasks that exist for one language. Ranking a roster by each gives two orderings, and the **pairwise inversion rate** is the fraction of model pairs they disagree about — 0% for identical orderings, about 50% for unrelated ones.

For example, `granite-311m` is globally 12/25 but falls to the local bottom quartile in 16 of 17 audit languages; `harrier-0.6b` is 3/25 and bottom-quartile in 10/17, and `Seed1.6-embed` 9/25 in 6/17. All three belong to a five-model cluster with this profile (§4.4): on the headline number alone they look like strong picks.

Throughout, *aggregated MTEB scores* denotes the MTEB(Multilingual) retrieval aggregate. We audit these retrieval rankings against language-specific rankings across 17 languages and 3 nested model rosters. Our contributions are:

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 398–429
November 6-10, 2026 ©2026 Association for Computational Linguistics

1. **Three-tier rank-inversion audit (§4.1).** Across 17 languages and three nested rosters, curated-tier median inversion is 33% with 11/17 languages FDR-significant (Benjamini and Hochberg, 1995); the extended-tier τ is inflated by roster composition (§4.4).

2. **Evidence consistent with asymmetric task coverage as the mechanism we test (§4.2, §4.3).** Three task-fixed probes localise disagreement to aggregation structure, and inversion rate correlates negatively with per-language task count (Spearman ρ=−0.59 curated, −0.67 extended), the structural prediction of an aggregation account.

3. **A cluster of hidden-failure models with a shared profile (§4.4).** Under the rank-based hidden-failure criterion, `granite-311m` affects 16/17 curated languages, `harrier-0.6b` 10/17, and three further models (`Seed1.6-embed` (ByteDance Seed Team, 2025), `inf-retriever-v1` (Yang et al., 2025), `granite-97m-r2`) appear in ≥ 3 languages. The cluster shows a notably shallower English–multilingual NDCG gap than non-cluster models, consistent with stronger relative English-task representation; we frame this as exploratory characterisation (§4.4).

4. **An actionable aggregation-side mitigation (§4.3, App. E).** A language-balanced aggregator (per-language mean first, then macro-average) cuts curated-tier median inversion from 33% to 14.5% (leave-one-out τ=0.71), the lowest among the held-out aggregators we test; we hold the target language out of its own predictor because the self-inclusive variant leaks it.

5. **Multilingual RAG benchmark and supporting downstream evidence (§5.1).** We release a six-configuration RAG benchmark (IT/JA/HI × finance/law); retrieval-side inversions propagate to answer-quality inversions under cross-vendor LLM judges and independent native-speaker annotation.

Scope of claims. We do not claim MTEB is broken or that aggregation is the sole driver of disagreement. Our contribution is an empirical audit — statistical validation, downstream evidence and an actionable aggregation-side mitigation — rather than a new evaluation methodology: the language-balanced aggregator is a macro-average, and we claim novelty for the cross-language audit and its findings, not for the estimator. The methodological template applies to any of the 122 BelebeleRetrieval (Bandarkar et al., 2024) language variants with sufficient evaluated models.

Artifacts and roadmap. Code and the RAG benchmark are released at `https://github.com/ArchitRastogi20/mteb-reliability-audit` and `https://huggingface.co/datasets/ArchitRastogi/multilingual-rag-finance-law`; App. W maps every result to its entry point. §2 positions the audit, §3 introduces the snapshot, rosters and benchmarks, §4 audits rank agreement and its causes, and §5 tests downstream answer quality.

2 Related Work

MTEB and multilingual evaluation. Muennighoff et al. (2023) introduced MTEB; Enevoldsen et al. (2025) extended it as MMTEB with substantially more multilingual tasks. Both frameworks expose language- and task-filtered leaderboard views; we study the cost of consulting the headline *aggregated* score, which remains the default user-facing surface. Most multilingual backbones we evaluate derive from XLM-R-style pretraining (Conneau et al., 2020) or language-agnostic sentence-embedding work (Feng et al., 2022). Language-specific MTEB derivatives already advocate per-language reporting: beyond Polish (Poświata et al., 2026) and Turkish (Baysan and Gungor, 2025), dedicated suites now exist for French (Ciancone et al., 2024), Persian (Zinvandi et al., 2025), Russian (Snegirev et al., 2025), Chinese (Xiao et al., 2024), Arabic (Bhatia et al., 2025), and Vietnamese (Pham et al., 2026), alongside Scandinavian (Enevoldsen et al., 2024), Japanese (Li et al., 2026), African (Uemura et al., 2026), and Southeast-Asian (Ponwitayarat et al., 2026) suites. Their proliferation signals demand for the per-language view the headline aggregate withholds. These benchmarks are complementary to our audit: they cover 6 of our 17 audit languages, 60% (21/35) of the curated-tier hidden-failure cells of §4.4 fall in languages outside them, and their voluntary rosters cover at most 10 of our 25 curated models each: no hidden-failure model meets the 50% full-benchmark coverage threshold on any available fork, with `inf-retriever-v1` the

one partial exception, covering all retrieval tasks of the French and Chinese forks (App. X). No prior work conducts a cross-language statistical audit of aggregation reliability on the default surface. MIRAGE-Bench (Thakur et al., 2025) evaluates multilingual RAG *generation* quality across 18 languages and is complementary to our retrieval-side audit.

Aggregation critiques and benchmark fragility. Colombo et al. (2022) show that aggregate rankings can obscure systematic *across-task-type* differences; Ethayarajh and Jurafsky (2020) show that averaged metrics miss deployment distinctions; Dehghani et al. (2021) document the "benchmark lottery" along the *task-subset* axis; Søgaard et al. (2021) make a similar point for split selection; Card et al. (2020) highlight under-powering of NLP comparisons. Rodriguez et al. (2021) propose IRT-based aggregation, orthogonal to ours. Peyrard et al. (2021) show that paired evaluation alters conclusions averaging obscures, Rofin et al. (2023) apply social-choice voting rules to multi-task benchmark ranking, and Ghosh et al. (2025) aggregate under incomplete measurements, the coverage-asymmetry setting from another angle; all three concern cross-system or cross-task mechanisms, not the cross-language axis. We extend this line along axes none of these treat together: (i) **cross-language** aggregation rather than cross-system, cross-task, or cross-split; (ii) a **three-tier roster sensitivity audit** (§4.1) that separates roster-driven inflation from genuine ranking disagreement; (iii) **three task-fixed controlled probes** (§4.2) plus a direct test of the asymmetric-coverage account (§4.3); and (iv) a **cluster of identifiable hidden-failure models** (§4.4) with a quantitatively distinctive English–multilingual profile on the predictor task suite.

Retrieval benchmarks and leaderboard integrity. BEIR (Thakur et al., 2021) established zero-shot cross-domain IR evaluation and showed that models generalise unevenly across task types, a finding we extend to cross-language generalisation. MIRACL (Zhang et al., 2023) and Mr. TyDi (Zhang et al., 2021) focus on cross-lingual retrieval but do not analyse aggregation effects on model selection; both inherit non-uniform per-language task availability, the same structural feature that predicts MMTEB inversions here. BelebeleRetrieval (Bandarkar et al., 2024) provides a single retrieval task across 122 language variants built on the FLORES-200 parallel corpus (NLLB Team et al., 2024); we use it together with monolingual WikipediaRetrievalMultilingual to construct a two-task controlled probe of aggregation structure (§4.2). The TREC IR tradition develops formal system-pair significance tests for ranking comparisons (Smucker et al., 2007); we combine these with Benjamini–Hochberg FDR control (Benjamini and Hochberg, 1995). RAGEval (Zhu et al., 2025) provides the methodology we adopt for our domain-specific benchmark, which we extend to Italian, Japanese, and Hindi finance and law. Hybrid learned sparse retrieval (Formal et al., 2021) and hybrid fusion (Cormack et al., 2009) are orthogonal mitigation route complementary to the per-language selection step our paper offers. An alternative explanation for leaderboard-vs.-deployment discrepancies is training/evaluation contamination (Lewis et al., 2021); we discuss it in Limitations.

Multilingual embedding models. Our roster (Table 12) spans retrieval-objective families derived from XLM-R-style backbones (Conneau et al., 2020; Feng et al., 2022; Zhang et al., 2024; Li et al., 2023) and language-specialised distillation (Sturua et al., 2025); full per-family citations in App. F.

3 Data and Methodology

3.1 MTEB Evaluation Setup

Figure 1 gives a high-level overview of the audit. We evaluate 17 paper-roster models spanning 33M to 8B parameters (full specifications in App. F) on language-specific MMTEB tasks for Italian, Japanese, and Hindi, on the multilingual RAG benchmark (§3.2), and on the deployment benchmark (§3.3). We use NDCG@10 (Järvelin and Kekäläinen, 2002) throughout. All 18 MMTEB(Multilingual) retrieval tasks report NDCG@10 natively in the public leaderboard CSV; no metric harmonisation or normalisation was applied.

Definitions and provenance. **MTEB-agg** is the model's MMTEB Mean(Task) retrieval score from the public leaderboard (`https://huggingface.co/spaces/mteb/leaderboard`; snapshot 2026-05; CSV released). MMTEB exposes both arithmetic Mean(Task), the public-leaderboard headline number, and Borda count (Enevoldsen et al., 2025); we target Mean(Task), with a Borda replication in App. K. The **lang-avg** is the mean NDCG@10

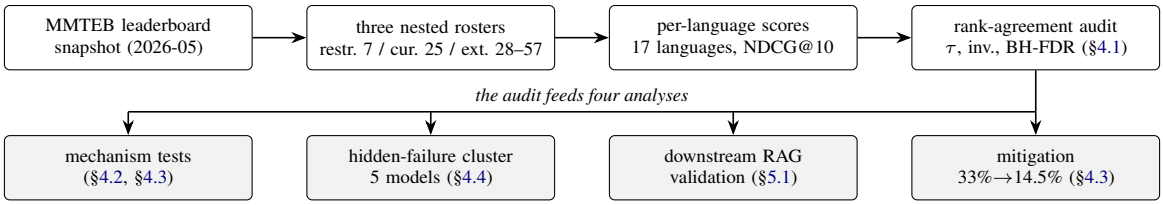


Figure 1: Overview of the audit. **Top row**: one frozen MMTEB leaderboard snapshot defines three nested model rosters, scored per language and compared against the aggregate ranking. **Bottom row**: the four analyses the audit feeds.

across the MTEB retrieval tasks available for that language. For the 17-language audit (Table 2) it is computed from the downloaded public per-language results, so predictor and outcome come from the same snapshot; per-language cells in Table 13 are own runs wherever own-run coverage exists and public per-language leaderboard values otherwise (marked in the caption). The **aggregation gap** Δ = lang-avg − MTEB-agg (positive means MTEB understates). A **hidden failure** is defined either heuristically (lang-avg$<$50 and MTEB-agg$>$75; stable under $\pm$5 pts) or under a principled threshold-free rank-based criterion (top-50% MTEB rank *and* bottom-25% lang rank; §4.4); the heuristic set is a strict subset of the rank-based set. A failure is *documented* iff a public model card or training-data report identifies a coverage gap consistent with the observed collapse, and *hypothesis-only* otherwise. We use the term at two scopes and label them throughout. **Cluster** hidden failures are the model–language cells of the five-model rank-based cluster read off leaderboard rankings in the curated and extended tiers (§4.4). **Case-study** hidden failures are those found in our own-run IT/JA/HI and Swahili analyses (`snowflake`, `nemotron-8b`; e.g. Table 17). The two use different rosters and score sources and name disjoint model sets here.

Three-tier model roster. How far the two rankings diverge depends on *which* models are ranked, and two properties of the roster confound that — its **size** and the **judgment** behind it — so we vary both across three nested tiers (Table 1; membership and ranks in App. B). *Restricted* holds size at the original paper roster; *extended* removes roster judgment by taking every public submission, but favours multilingual-first models that opt into broad evaluation (§4.1); *curated* sits between and is our headline, approximating the pool a practitioner chooses from.

The curated tier is the restricted tier plus 18 models we selected *by hand* as deployment-relevant at the 2026-05 snapshot: all have $\geq$ 2-language MMTEB(Multilingual) retrieval coverage, and each was admitted on a multilingual-retrieval training or instruction-tuning claim, high MMTEB retrieval standing, or visible practitioner uptake. No mechanical rule derives membership and we claim none; what makes the tier exactly auditable is that the 25-model list is frozen and released in full (`analysis/roster_membership.csv`). Three guards carry the argument instead: alternate re-derivations (Table 7), the judgment-free extended tier, and an independent later snapshot (App. Z).

Tier	n/lang	Selection criterion	Purpose
Restricted	6–7	paper roster (App. F) ∩ MMTEB(Multilingual) Retrieval submissions	same-pipeline core at original size
Curated	24–25	restricted + 18 deployment-relevant models selected by hand and released as a frozen list (§3.1)	deployment-realistic headline tier
Extended	28–57	all public MMTEB(Multilingual) retrieval submissions per language (judgment-free)	maximal power; removes roster judgment

Table 1: The three nested model rosters, which vary roster size and roster judgment independently (§3.1). *Curated*, the headline tier, is a hand-selected, frozen, released list, not the output of a mechanical rule; *extended* carries no roster judgment but favours multilingual-first broad submitters. Per-language n variation is detailed in the Table 2 caption; extended ranges from n=28 (spa) to n=57 (rus). Other named model subsets used in local prose are collected in Table 8.

Other named subsets. Local prose elsewhere also names several narrower subsets, collected in Table 8 (App. B): the *paper roster* (n=17, App. F); the *legacy IT/JA/HI subset* (n=13–15, App. A); the *deployment* set (n=15 core +2 comparators, §3.3);

and two task-fixed-probe subsets (n=14 excluding `nemotron-8b`, n=12 for IT–HI Wikipedia, §4.2).

Per-language task coverage. Of the 18 MMTEB retrieval tasks (11 English-derived, 7 multilingual; App. C), each language's lang-avg uses only the 2–4 exposed for it (median 2; n_t, Table 2); this coverage asymmetry is the mechanism we test in §4.3.

3.2 Multilingual RAG Benchmark

We introduce a unified Multilingual RAG Benchmark over six configurations, {Italian, Japanese, Hindi} × {finance, law}, written *it-fin*, *it-law*, *ja-fin*, *ja-law*, *hi-fin* and *hi-law*. Each contains 250 documents and the 375–472 generated queries (mean 444, of up to 500) that pass all validation gates; all 17 models are evaluated on all six configurations under an identical protocol. Italian, Japanese, and Hindi were chosen to span scripts and resource levels (Latin/CJK/Indic) and MTEB task-coverage levels (n_t=2/2/4), subject to the cost of dataset construction and independent native-speaker validation; the released pipeline extends to any BelebeleRetrieval language.

Unified construction pipeline. Following RAGEval (Zhu et al., 2025), all six benchmarks were generated by a 5-stage GPT-5.4-Nano/Mini pipeline (OpenAI, 2026) (config → article → queries → reference refinement → RAG answer+keypoints) with LLM ground-truth gating at every stage. Averaged over the three *answerable* query types only (factual, multi-hop and summarisation; both domains, unanswerable excluded), the mean Stage-3–5 pass rate is JA 95.1%, IT 89.7%, HI 81.2%; the largest shortfalls are hi/finance multi-hop (59%) and summarisation (49%), then it/law summarisation (77%) and it/finance multi-hop (82%). Unanswerable-confirmation rates are given per cell in Table 15. All GPT-5.4-Nano/Mini pipeline calls used the API defaults (`temperature=1.0`, `top_p=1.0`) with Pydantic structured outputs; the cross-judge (§5.1) and closed-book baseline (App. U) used `temperature=0`.

Native-speaker validation. 50 factual + 50 multi-hop queries per language were spot-verified by two annotators each; rating-level accuracy ≥ 92% on all six cells.

BM25 baselines and tokenisation. BM25 (Robertson and Zaragoza, 2009) is a reference baseline, not a comparator in the audit: our claims concern disagreement *among dense models*, and the downstream evidence is retriever-versus-retriever. On five of six configurations (it-fin, it-law, ja-law, hi-fin, hi-law) BM25 leads or ties the best dense retriever and is well behind only on ja-fin (Table 5); hi-fin (92.3) is a Devanagari tokenisation artifact and is excluded from headline dense comparisons. Tokenisation is case-folded whitespace for Italian and Hindi and character-level for Japanese; a case-sensitive Italian tokeniser instead collapses it-fin BM25 to 2.1, a caveat for Latin-script BM25 baselines rather than a property of the benchmark.

3.3 Deployment Benchmark

We benchmark query throughput (QPS), query-time VRAM, and p50/p99 latency on a single NVIDIA RTX 3090 (24 GB) for the 15-model intersection of the paper roster with models fitting the 24 GB budget, plus two deployment-relevant comparators (`jina-v3`, `pplx-embed-v1-0.6b`); `Qwen3-8B` appears in Table 19 with n/a entries flagging its index-build OOM, and `nemotron-8b` was not benchmarked on this hardware. Each model is run for 3 trials with 1,000 corpus documents and 1,744 queries; means reported. **QPS is embedding-only throughput** (encoding queries against a pre-built index); end-to-end RAG latency (retrieval + LLM generation) is dominated by the generator and not measured here.

4 Results: Aggregation Bias

4.1 Three-Tier Rank-Inversion Audit

Table 2 reports per-language Kendall τ and pairwise inversion rate at all three roster sizes for 17 languages. The headline pattern is non-monotonic: median τ is 0.43 in the restricted tier, drops to **0.34** in the curated tier (median inversion 33%, 11/17 languages reject H_0 after BH FDR), and rises to 0.69 in the extended tier.

Both sides of this table come from one snapshot. The per-language lang-avg is computed from the downloaded public per-language MMTEB results, not our own runs, so an inversion here cannot be an artifact of pairing a leaderboard predictor with an own-run outcome. Own runs enter only in the IT/JA/HI case studies, the Swahili replication and the RAG benchmark, as same-pipeline

corroboration (App. T).

The extended-tier rise is a composition effect, not signal. Random size-25 subsamples of the extended tier give median τ=0.70, indistinguishable from the full extended-tier 0.69 and far above the curated 0.34 at matched n. The curated 25 is a strict subset of the extended roster, but the hidden-failure cluster (§4.4) is diluted in it: 4 of 25 curated models (16%) against 5 of 57 extended (9%), so a size-25 extended draw holds ≈2.2 of them on average. The extended roster is additionally dominated by multilingual-first submitters whose per-language and global rankings agree, whereas the curated tier concentrates the deployment-relevant English-strong models (e.g., `KaLM-Embedding-Gemma3-12B` (Zhao et al., 2026), `granite-311m`) that surface the disagreement, and is the practitioner-realistic estimate.

The 33% inversion rate, not τ, is the substantive measure; the τ CIs of App. A use 10,000 model-level resamples (Efron and Tibshirani, 1993).

Top-1 selection regret. The MTEB-aggregate top-1 is `KaLM-Embedding-Gemma3-12B` for every language; the per-language oracle is `Qwen3-8B` (13/17) or `nemotron-8b` (Babakhin et al., 2025) (4/17). Top-1 regret averages 3.3 NDCG@10 (range 0.1–5.5); the four highest-regret languages (jpn, kor, rus, fra) all have n_t=2, consistent with coverage asymmetry (§4.3). The cost is concentrated at top-1: a top-3 shortlist removes almost all of it (mean regret 0.07 NDCG@10, exactly zero in 13 of 17 languages).

Calibration against random. Two random rankings of n=25 items have an expected inversion rate of 50% (10,000 Monte Carlo draws, 95% CI $[36.0\%, 64.3\%]$). The curated 33% sits below the 1st percentile of this null and the extended 15.4% below all draws, so the thin per-language averages still carry strong shared signal.

4.2 Task-Fixed Probes: Localising Disagreement to Coverage Asymmetry

If coverage asymmetry drives disagreement, cross-language agreement should collapse to high values once the task is held constant. We test this on the IT/JA/HI subset (task counts 2/2/4) with **three** shared retrieval anchors spanning parallel-corpus (Belebele), monolingual (Wikipedia), and cross-lingual (MIRACL) construction (Table 3); using all three argues against parallel-corpus neutrality as the driver.

Construction. For each language pair we compare Spearman ρ between the two languages' per-model *aggregation gaps* (task or lang-avg score − MTEB-agg) on the shared task against ρ on the full per-language averages; higher shared-task ρ points to task mixture rather than model variance. All ρ use the n=14 paper roster (excluding `nemotron-8b`, footnote † of Table 6); IT–HI Wikipedia drops two models lacking Wikipedia scores (n=12).

Holding the task constant lifts cross-language agreement (Table 3). The Italian–Japanese lift on Belebele is significant under a paired bootstrap on $\Delta = \rho_{\text{full}} - \rho_{\text{task}}$ (B=10,000); the Italian–Hindi Wikipedia lift points the same way but is underpowered at n=12; Japanese–Hindi, already high under the full averages, does not move on MIRACL. The monolingual Wikipedia control *exceeds* the parallel-corpus Belebele on the same pair, so parallel-corpus neutrality does not drive the high task-fixed agreement (per-model gaps visualised in Figure 5, App. G). Per-task model counts (Table 21) reflect the subset with a non-missing NDCG@10 on the specific task.

A saturation alternative is not supported by per-task variance (App. O). We also reran the headline rank-correlation under MMTEB's Borda-count rank aggregator (App. K): IT/JA/HI inversion rates rise to 31–41% vs. 20–27% under Mean(Task), ruling out arithmetic averaging as the driver. §4.3 tests the structural prediction directly across all 17 audit languages.

4.3 Task Coverage Predicts Inversion Rate

The structural setup is per-language inhomogeneous: each language's lang-avg uses a different, non-identical subset of the retrieval tasks exposed for it (2–4 per language; §3.1), and MTEB-agg additionally contains 11 English-derived tasks not in any lang-avg. This predicts (a) higher inversion for languages with fewer exposed tasks (*symptom*), and (b) that restricting MTEB-rank to multilingual tasks should beat coverage reweighting (*localisation*: the disagreement should track English-task content rather than coverage breadth); both hold below. For each tier, we compute Spearman ρ between n_t and the per-language inversion rate across all 17 audit languages. With 17 language points and $n_t \in \{2, 3, 4\}$ the design cannot separate coverage from language identity, script or task difficulty, so we read what follows as a consistency check on an aggregation account rather than a causal identifica-

ISO	Lang.	n_t	**Restricted** (n=7) τ	inv%	**Curated** ($n\sim$25) τ	inv%	**Extended** (n=28–57) τ	inv%
fra	French	2	0.52	23.8	0.25	37.3	**0.64**	18.2
ind	Indonesian	2	0.52	23.8	0.26	37.0	**0.68**	15.8
tel	Telugu	2	0.33	33.3	0.26	37.0	**0.65**	17.3
kor	Korean	2	0.43	28.6	0.28	36.0	**0.69**	15.3
swa	Swahili	2	0.24	38.1	0.30	35.0	**0.66**	17.1
fas	Persian	3	0.20	40.0	0.30	34.8	**0.69**	15.4
tha	Thai	2	0.43	28.6	**0.32**	34.0	**0.68**	16.0
jpn	Japanese	2	0.43	28.6	**0.33**	33.3	**0.69**	15.5
rus	Russian	2	0.52	23.8	**0.34**	33.0	**0.65**	17.3
ita	Italian	2	0.20	40.0	**0.36**	31.9	**0.72**	14.1
zho	Chinese	3	0.52	23.8	**0.39**	30.3	**0.69**	15.4
deu	German	4	0.33	33.3	**0.41**	29.3	**0.74**	13.2
ben	Bengali	3	0.20	40.0	**0.42**	29.0	**0.72**	13.9
ara	Arabic	3	0.52	23.8	**0.45**	27.3	**0.73**	13.3
hin	Hindi	4	0.33	33.3	**0.46**	26.8	**0.74**	13.1
spa	Spanish	3/3/4[§]	0.62	19.0	**0.49**	25.7	**0.92**	4.0
vie	Vietnamese	2	0.52	23.8	**0.59**	20.7	**0.75**	12.6
Median			0.43	28.6	0.34	33.0	0.69	15.4
Reject H_0 after BH FDR				0/17		11/17		17/17

Table 2: Three-tier sensitivity audit of MTEB(Multilingual) ranking agreement across 17 languages. n_t = number of MTEB retrieval tasks exposed for the language. τ = Kendall rank correlation between MTEB-aggregate and language-specific lang-avg; inv% = pairwise inversion rate. Bold = rejects H_0 at q=0.05 after Benjamini–Hochberg FDR correction within tier. Rows sorted by curated τ ascending. *Restricted-tier per-language values are noisy at n=7 and not comparable across languages; we report them for tier-internal trends only, the median (28.6%) being the substantive restricted-tier summary.* The extended tier's high τ partly reflects roster composition: the hidden-failure cluster (§4.4) is diluted in the broader roster, which is dominated by multilingual-first submitters. [§] Spanish n_t is the only tier-varying value: 3 retrieval tasks (Belebele, MIRACL, MLQA) in the restricted and curated tiers, 4 in the extended tier (XPQARetrieval, exposed only to the broader roster); the Figure 3 extended-tier ρ=−0.67 uses n_t=4 for Spanish, the other 16 languages being constant across tiers. In five languages (`ita`, `hin`, `ben`, `deu`, `fas`) `ml-e5-small` lacks MMTEB results, giving n=6/24 rather than 7/25 and hence the non-21-pair restricted-tier percentages (e.g., 40.0%=6/15).

Pair	Anchor	n	ρ_{full}	ρ_{task}	Δ [95% CI]	p_Δ
IT–JA	Belebele	14	0.512	**0.938**	−0.426 [−0.97, −0.07]	0.004
IT–HI	Wikipedia	12	0.594	**0.937**	−0.343 [−1.03, 0.05]	0.118
JA–HI	MIRACL	14	0.829	**0.837**	−0.009	0.963

Table 3: Task-fixed probes. ρ_{full}: Spearman ρ between the two languages' per-model *aggregation gaps* (lang-avg score − MTEB-agg) under full per-language averages; ρ_{task}: same with the task fixed on the shared anchor (Wikipedia ρ=0.937: p<0.001); Δ: paired bootstrap, B=10,000. A negative Δ means the orderings agree *better* once the task is held fixed, as an aggregation account predicts; n counts models scored on both sides. A secondary IT–HI Belebele anchor gives ρ_{task}=0.662 (n=14).

tion.

The prediction holds at deployment-realistic and broad roster sizes: Spearman $\rho = -0.59$ (95% CI $[-0.85, -0.18]$) in the curated tier and $\rho = -0.67$ (95% CI $[-0.92, -0.28]$) in the extended tier (Figure 3, App. D). Because n_t takes only three distinct values, ties are extensive; a tie-aware Kendall τ_b with a 10,000-draw permutation test agrees (τ_b=−0.481, p=0.017 curated; τ_b=−0.587, p=0.002 extended). The restricted tier is too small (n=7 models) to recover this signal.

Multilingual-only aggregator. Table 11 (App. E) ranks six aggregators on the curated tier. Restricting MTEB-rank to the 7 multilingual predictor tasks lowers median inversion below the Mean(18) headline and raises Kendall τ; the per-language drops are largest for Arabic (27→10), Hindi (27→11) and Bengali (29→13), and Swahili, a sparse-coverage language, is the only one it leaves unchanged (App. D). Restricting instead to the 11 English-derived tasks moves both metrics the wrong way, as does a $1/n_t$ reweighting that down-weights broad-coverage tasks, localising the issue to English-task content rather than coverage breadth. A language-balanced alternative (per-language mean first, then macro-average) is the strongest of the six, dropping median inversion to **14.5%** (LOO, τ=0.71; 17/17 languages

Model	Best global	Med. local	#Lang (cur.)	MTEB agg
granite-311m	12	24	16/17	0.642
harrier-0.6b	3	21	10/17	0.701
Seed1.6-embed	9	20	6/17	0.653
inf-retriever-v1	7	20	3/17	0.658
granite-97m-r2	28	40.5	4/17*	0.595

Table 4: Hidden-failure cluster (curated tier, n=25). **Best global**: best (lowest) global rank in MTEB(Multilingual) Retrieval. **Med. local**: median local rank across affected languages (local pool size n=24 or 25 in curated). **MTEB agg**: MMTEB Mean(Task) retrieval score on $[0, 1]$. *`granite-97m-r2` appears as a hidden failure in the extended tier only (4/17 languages, applying the bottom-25% threshold within the n≈53–57 extended pool); it is absent from the curated 25-model roster and so is never classified there. The four curated-tier models account for the **35 curated-tier hidden-failure cells** (16+10+6+3; $\geq$ 3-language criterion); `granite-97m-r2` adds 4 extended-tier cells.

improve, paired CIs in App. E) and generalising to the judgment-free extended tier.

The aggregator was found by a sweep. We did not pre-specify it: the first aggregator experiment compared only Mean(18), Mean(7) and Mean(11), and the language-macro variant entered hours later beside the $1/n_t$ reweighting that fails. The three guards above — leave-one-out evaluation, 17-of-17 paired improvement, and extended-tier replication — are what stand against over-fitting to the curated tier.

4.4 A Cluster of Hidden Failures

A model–language pair is a *cluster hidden failure* (§3.1) under our threshold-free rank-based criterion: the model ranks in the **top half globally** on MTEB(Multilingual) Retrieval but in the **bottom quartile locally**. This criterion strictly contains the heuristic 50/75 NDCG cutoff and remains stable under alternative thresholds (App. L). Across the curated and extended tiers we identify five models that satisfy the criterion in $\geq$ 3 languages. Four are in the curated tier (`granite-311m`, `harrier-0.6b`, `Seed1.6-embed`, `inf-retriever-v1`); `granite-97m-r2` is absent from the curated 25-model roster and surfaces only at the extended-tier population (Table 4).

Exploratory profile: shallower English–multilingual gap. We classify the 18 retrieval tasks as predominantly English-derived (11 tasks) or multilingual (7 tasks; full taxonomy in App. C) and compute the per-model gap $\Delta = \overline{\text{eng}} - \overline{\text{multi}}$ (per-model values in Table 9, App. C). Cluster median $\Delta = -0.026$ vs. non-cluster median -0.120: the cluster sits 0.094 NDCG higher.[1] The only two positive deltas across the 26 evaluated models (`harrier-0.6b`, `granite-97m-r2`) are both cluster members. Most curated-tier models score higher on the multilingual subset (which is broader and less specialised); the cluster shows a notably shallower gap, consistent with stronger relative English-task representation. `granite-311m` is the canonical case: globally rank 12/25, local rank $\geq$ 21 on 16 of 17 audit languages (ranks in App. B); exception: Swahili, rank 16.

Same-pipeline replication (case-study scope). The *case-study* pattern (`snowflake`, `nemotron-8b`) reproduces under own-run-vs.-own-run ranking in three settings (§4.2, App. Q, App. R), and per-cell public-vs.-own-run gaps concentrate on those cells (App. T). The cluster above is leaderboard-derived on both sides, so no pipeline mix arises there.

No cluster member is a hidden failure in *both* tiers (§4.1), and the MMTEB language filter does not resolve this, since each underlying task is itself a cross-language average.

5 Results: Downstream RAG Evidence

5.1 Downstream Evidence: Inversions Propagate to RAG

Two tests probe whether the retrieval-side inversions surface downstream on the six-configuration benchmark of §3.2:

Cross-vendor LLM judges agree on RAG inversions. We retrieve top-3 with each retriever and score keypoint coverage under two judges from different vendors (GPT-5.4-Mini and Claude-Haiku-4.5 (Anthropic, 2025); cf. ARES (Saad-Falcon et al., 2024) for learned RAG evaluation, where we instead score against keypoints fixed at generation time). Both judges put `bge-m3` ahead of `snowflake` on ja-fin by the same keypoint-coverage margin, and each is significant on its own

[1] Mann–Whitney one-sided U=91, p=0.005. The cluster is defined from per-language ranking data, so this contrast is not an independent test of membership; we report the p-value descriptively.

Model	P	Italian fin	Italian law	Japanese fin	Japanese law	Hindi fin	Hindi law
BM25	n/a	78.7	62.9	50.3	31.0	92.3	79.0
ml-e5-small	118M	66.0	42.7	83.8	19.2	71.2	58.0
jina-v5-nano	212M	**78.3**	52.5	96.3	24.6	73.5	58.8
bge-m3	568M	72.8	47.7	93.0	23.4	78.1	61.1
snowflake	568M	64.5	37.1	75.3	12.0	77.2	44.8
nemotron-8b	7.5B	44.0*	30.9	88.2	24.8	76.0	41.9
Qwen3-4B	4B	77.0	**59.0**	95.9	**28.6**	**81.3**	61.7
Qwen3-8B	8B	74.4	55.8	**96.9**	27.4	78.5	**62.8**
text-emb-3-small	API	70.6	48.3	91.3	22.0	31.2	18.7
text-emb-3-large	API	67.4	47.4	85.4	21.7	57.3	34.2
cohere-v3	API	69.7	43.7	91.7	25.4	73.9	57.7
gemini-emb-2	API	60.8	19.8	28.6	17.9	68.5	25.3

Table 5: Headline RAG NDCG@10 (representative subset; all 17 roster models + 4 API endpoints in `analysis/analysis_output/unified_rag_results_v2.csv`). All values are NDCG@10 × 100; bold = best neural per column; tokenisation: §3.2. API endpoints: `text-embedding-3-small/-large` (OpenAI, 2024), `cohere-embed-multilingual-v3` (Cohere, 2023), `gemini-embedding-2` (Google, 2026). * `nemotron-8b` on it-fin reproduces the MTEB Italian hidden failure (§3.1): competitive on JA/HI but collapses on Italian.

under a paired Wilcoxon signed-rank test (Table 14, App. H), ruling out a same-family-judge artifact (Zheng et al., 2023).

Native-speaker annotation: Hindi corroborates, Italian is null. Independent native-speaker annotation on hi-fin and it-fin (n=50/50, App. V) gives Δ_{human}=0.234 on Hindi (p=0.001, inter-annotator $\rho \geq 0.94$); the it-fin contrast is null (p=0.88) despite a 93.0 vs. 65.3 MTEB Italian-avg gap (Table 13), so retrieval-side inversions frequently but not universally propagate. The pair's RAG retrieval gap (78.3 vs. 64.5, Table 5) is smaller than its MTEB gap and may be partly absorbed by the generator. Hidden failures (§4.4) still reproduce in RAG: `nemotron-8b` on it-fin collapses while remaining competitive on JA/HI (Table 5; Pareto frontiers in Figure 7, App. I).

Closed-book baseline confirms retrieval is value-adding. GPT-5.4-Mini with no retrieved context (App. U, n=2,425) recovers only 14.5–35.8% of gold keypoints — query-weighted mean 21.3%, and strictly below the retrieval-augmented signal on every one of the six configurations (Table 28) — so the contrasts above are not driven by parametric memory. Per-config MTEB-to-RAG ρ is weaker on both Italian configurations than on any Japanese or Hindi one (Table 14; App. H), so retrieval-step ranking is a less reliable downstream signal in Italian regardless of metric.

Practitioner recipe. Prefer per-language MMTEB scores over the aggregate; failing that, a BelebeleRetrieval $\geq$ 60 pre-filter plus smallest-Pareto selection avoids all four IT/JA/HI case-study hidden failures in our 17-model sample — three documented, one hypothesis-only (App. P).

6 Conclusion

At deployment-realistic roster size, aggregated MTEB(Multilingual) retrieval scores disagree with per-language rankings on a median 33% of model pairs (11/17 languages FDR-significant); the disagreement is predicted by asymmetric task coverage, concentrates in an identifiable hidden-failure cluster, and propagates to downstream RAG quality under cross-vendor judges and native-speaker annotation. The most actionable mitigation is aggregation-side: a language-balanced macro-average cuts median inversion to 14.5% (leave-one-out τ=0.71).

Three implications follow. A **benchmark maintainer** can fix most of this at the display layer, by default-displaying the language-balanced aggregate or the per-language view: the inversions come from pooling tasks with unequal language coverage, not from the scores themselves — hence they persist under Borda (App. K) but largely vanish once languages are balanced. A **practitioner** should read per-language scores, or else apply the screen above. Two questions remain open: *why* the cluster has this profile, and whether inversions hold beyond retrieval and the classification and STS replicated in App. Y; clustering and reranking are left unaudited.

Limitations

Retrieval-specific scope. The audit is of the *retrieval* category of MMTEB(Multilingual). The classification and STS replication in App. Y is leaderboard-only, with no own runs; clustering is not audited, and reranking cannot be audited at all: it exposes fewer than two tasks for every one of our audit languages, itself an instance of the coverage asymmetry we study. Nothing here is a claim about MTEB's non-retrieval categories.

Sparse per-language task coverage. Within retrieval, the audit covers the 17 languages with

≥ 2 MTEB retrieval tasks; substantive claims rest on the joint pattern across rosters and the FDR-corrected curated-tier rejections (11/17). 10 of 17 languages have only n_t=2, so the local lang-avg is itself a 2-point average susceptible to per-task idiosyncrasies. Noise of this kind attenuates rank correlations toward zero and cannot manufacture the negative coverage association of §4.3 or the null-calibration margin of §4.1, so the curated-tier estimates are, if anything, conservative; the language-macro aggregator (App. E) is a more robust target where coverage permits.

Limited downstream validation. The downstream RAG evidence covers 3 of the 17 audit languages (IT/JA/HI), chosen to span scripts, resource levels and task coverage (§3.2), across two domains (finance and law). We therefore do not show that retrieval-side inversions propagate in the other 14 audit languages, in other domains, or at deployment corpus sizes, and even within these three languages propagation is not uniform: the Italian contrast is null downstream (§5.1).

Roster judgment and aggregator selection. Two choices in the paper are ours rather than the data's. The curated tier is a hand-selected, judgment-based roster (§3.1); we release the frozen 25-model list so the tier is exactly auditable, and report the judgment-free extended tier and three alternate re-derivations beside it. The language-balanced aggregator was likewise found by an exploratory sweep of alternatives rather than pre-specified (§4.3).

Hidden-failure attribution. The cluster's shallower English–multilingual gap is consistent with stronger relative English-task representation, but causal attribution to training-data distribution requires statistics several cluster-model authors do not publish; we frame the contrast as exploratory (§4.4). Smaller-effect cases below the threshold-free criterion may exist.

Synthetic benchmark and judge–generator overlap. The RAG benchmark and primary judge share the GPT-5.4 family (Zheng et al., 2023); we mitigate this with a Claude-Haiku-4.5 cross-judge (§5.1) and native-speaker annotation on Hindi and Italian (App. V). An 8-gram contamination check against language-matched Wikipedia and BelebeleRetrieval gives overlaps of 0.63% / 0.04% respectively (word-level for IT/HI, character-level for JA), averaged over the six (language, domain) configurations.

RAG benchmark scale. Each configuration uses a 250-document corpus; top-10 retrieval covers 4% of the pool, generous compared to deployment scales (10^4–10^6 documents). The Italian-side human null (§5.1) may partly reflect retriever robustness at this corpus size rather than absent downstream propagation; deployment-realistic effects likely exceed our conservative estimates.

Models and deployment scope. The historical multilingual baselines XLM-R, LaBSE, and mGTE are positioned in Related Work as backbone context but are not part of the 17-model rank-inversion roster. The deployment benchmark intentionally targets the single-GPU commodity self-hosting setting that leaderboard-driven selection most often serves (RTX 3090, 1,000-document corpus; App. N); API endpoints are compared on retrieval quality (Table 5) but not on serving latency, and large-scale production serving is out of scope.

Ethical Considerations

This work evaluates publicly available models and datasets; synthetic benchmark content contains no personally identifiable information. Native-speaker validation (§3.2, App. V) used volunteer annotators with graduate-level proficiency, who consented to release of anonymised aggregated scores; protocol and raw annotations are released. We follow the ACL Code of Ethics. GPT-5.4 (OpenAI, 2026) was used for benchmark generation (§3.2) and grammar correction; no claims or analyses were AI-introduced.

References

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Compact and robust text embedding models using task-targeted distillation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pages 4454–4458. Association for Computing Machinery. ArXiv:2602.15547.

Anthropic. 2025. Claude Haiku 4.5. Anthropic Model Documentation. Snapshot used: `claude-haiku-4-5-20251001` (released 2025-10-15), accessed via the Anthropic Messages API on 2026-05-01.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. Granite embedding models. *Preprint*, arXiv:2502.20204.

Parul Awasthy, Aashka Trivedi, Yushu Yang, Ken Barker, Yulong Li, Bhavani Iyer, Martin Franz, Juergen Bross, Meet Doshi, Vignesh P, Vishwajeet Kumar, Todd Ward, Abraham Daniels, Madison Lee, Luis Lastras, Jaydeep Sen, and Radu Florian. 2026. Granite embedding multilingual R2 models. *Preprint*, arXiv:2605.13521.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. 2025. Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4669–4685, Albuquerque, New Mexico. Association for Computational Linguistics.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2020. AILA 2019 precedent & statute retrieval task. Zenodo dataset, doi:10.5281/zenodo.4063986. Dataset of the FIRE 2019 AILA track; overview in CEUR Workshop Proceedings Vol. 2517, pp. 1–12.

Nicolas Boizard, Théo Deschamps-Berger, Hippolyte Gisserot-Boukhlef, Céline Hudelot, and Pierre Colombo. 2026. BidirLM: From text to omnimodal bidirectional encoders by adapting and composing causal LLMs. *Preprint*, arXiv:2604.02045.

ByteDance Seed Team. 2025. Built on Seed1.6-Flash, Seed-1.6-Embedding launched. ByteDance Seed Blog. Multimodal embedding model supporting text, image, and video; accessed 2026-05-12.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. MTEB-French: Resources for French sentence embedding evaluation and analysis. *Preprint*, arXiv:2405.20468.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Cohere. 2023. Introducing Embed v3. Cohere Blog. Accessed: 2026-05-01.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. 2022. What are the best systems? new perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, New York, NY, USA. Association for Computing Machinery.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *Preprint*, arXiv:2107.07002.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The Scandinavian Embedding Benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, volume 37, pages 40336–40358. Curran Associates, Inc.

Sedigheh Eslami, Maksim Gaiduk, Markus Krimmel, Louis Milliken, Bo Wang, and Denis Bykov. 2026. Diffusion-pretrained dense and contextual embeddings. *Preprint*, arXiv:2602.11151.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, New York, NY, USA. Association for Computing Machinery.

Adhiraj Ghosh, Sebastian Dziadzio, Ameya Prabhu, Vishaal Udandarao, Samuel Albanie, and Matthias Bethge. 2025. ONEBench to test them all: Sample-level benchmarking over open-ended capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32445–32481, Vienna, Austria. Association for Computational Linguistics.

Google. 2026. Gemini embedding 2: Our first natively multimodal embedding model. Google Blog. Model ID: `gemini-embedding-2-preview`; released March 2026, accessed 2026-05-01.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Søren Vejlgaard Holm. 2024. Are GLLMs danoliterate? Benchmarking generative NLP in Danish. Master's thesis, Technical University of Denmark. DTU Compute.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-LLM collaborative dataset for generative information-seeking with attribution. *Preprint*, arXiv:2307.16883.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. 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*, pages 1000–1008, Online. Association for Computational Linguistics.

Shengzhe Li, Masaya Ohagi, Ryokan Ri, Akihiko Fukuchi, Tomohide Shibata, and Daisuke Kawahara. 2026. JMTEB and JMTEB-lite: Japanese massive text embedding benchmark and its lightweight version. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 7423–7434, Palma de Mallorca, Spain. ELRA Language Resource Association.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-CPR: A multi domain Chinese dataset for passage retrieval.

In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 3046–3056. Association for Computing Machinery.

Xing Han Lu, Siva Reddy, and Harm de Vries. 2023. The StatCan dialogue dataset: Retrieving data tables through conversations with genuine intents. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2799–2829, Dubrovnik, Croatia. Association for Computational Linguistics.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. SFR-Embedding-Mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Microsoft. 2026. harrier-oss-v1: A family of multilingual text embedding models. Hugging Face Model Repository. 0.6B parameter multilingual embedding model; decoder-only architecture with last-token pooling; 94 languages; 32K context; accessed 2026-05-12.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv:2402.09906.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, and 1 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846. Introduces the FLORES-200 parallel corpus; arXiv:2207.04672.

OpenAI. 2024. New embedding models and API updates. OpenAI Blog. Introduces `text-embedding-3-small` and `text-embedding-3-large`; accessed 2026-05-01.

OpenAI. 2026. GPT-5.4 model family. OpenAI API documentation and launch announcement. Snapshots used: `gpt-5.4-mini-2026-03-17` and `gpt-5.4-nano-2026-03-17` (released 2026-03-17), accessed via the Chat Completions API; model listing verified via `client.models.list()` on 2026-05-01. Launch announcement: https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315, Online. Association for Computational Linguistics.

Loc Pham, Tung Luu, Thu Vo, Minh Nguyen, and Viet Hoang. 2026. VN-MTEB: Vietnamese massive text embedding benchmark. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1705–1725, Rabat, Morocco. Association for Computational Linguistics.

Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Raymond Ng, Jann Railey Montalan, Thura Aung, Jian Gang Ngui, Yosephine Susanto, William Chandra Tjhi, Panuthep Tasawong, Erik Cambria, Ekapol Chuangsuwanich, and Sarana Nutanong. 2026. SEA-BED: How do embedding models represent Southeast Asian languages? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8788–8822, San Diego, California, United States. Association for Computational Linguistics.

Rafał Poświata, Sławomir Dadas, and Michał Wiktor Perełkiewicz. 2026. PL-MTEB: Polish massive text embedding benchmark. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 35601–35619, San Diego, California, United States. Association for Computational Linguistics.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *Journal of Biomedical Informatics*, 121:103865.

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

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. 2023. Vote'n'rank: Revision of benchmarking with social choice theory. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 670–686, Dubrovnik, Croatia. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pages 623–632, New York, NY, USA. Association for Computing Machinery.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Aleksandr Abramov. 2025. The Russian-focused embedders exploration: ruMTEB benchmark and Russian embedding model design. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 236–254, Albuquerque, New Mexico. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2025. Jina embeddings V3: Multilingual text encoder with low-rank adaptations. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, volume 15576 of *Lecture Notes in Computer Science*, pages 123–129. Springer Nature Switzerland. ArXiv:2409.10173.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad. 2025. MIRAGE-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 274–298, Albuquerque, New Mexico. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kosei Uemura, Miaoran Zhang, and David Ifeoluwa Adelani. 2026. AfriMTEB and AfriE5: Benchmarking and adapting text embedding models for African languages. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3697–3717, Rabat, Morocco. Association for Computational Linguistics.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 24 others. 2025. EmbeddingGemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

Voyage AI. 2025. voyage-3.5 and voyage-3.5-lite: Improved quality for a new retrieval frontier. Voyage AI Blog. General-purpose multilingual embedding model; 32K context; Matryoshka + quantization-aware training; released May 2025; accessed 2026-05-12.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 641–649, Washington DC, USA. Association for Computing Machinery. Introduces the C-MTEB Chinese benchmark; arXiv:2309.07597.

Junhan Yang, Jiahe Wan, Yichen Yao, Wei Chu, Yinghui Xu, Emma Wang, and Yuan Qi. 2025. inf-retriever-v1 (revision 5f469d7).

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2025. Arctic-embed 2.0: Multilingual retrieval without compromise. In *Second Conference on Language Modeling (COLM)*. ArXiv:2412.04506.

ZeroEntropy. 2026. Introducing zembed-1: The world's best text-embedding model. ZeroEntropy Blog. 4B parameter multilingual embedding model based on Qwen3-4B; zELO distillation from zerank-2 reranker; released March 2026; accessed 2026-05-12.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Wenbo Yang, Daiting Shi, and Xueqi Cheng. 2026a. Bagging-based model merging for robust general text embeddings. *Preprint*, arXiv:2602.05787.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. 2026b. F2llm-v2: Inclusive, performant, and efficient embeddings for a multilingual world. *Preprint*, arXiv:2603.19223.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, zhenyu liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2026. KaLM-embedding-v2: Superior training techniques and data inspire a versatile embedding model. In *The Fourteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. RAGEval: Scenario specific RAG evaluation dataset generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. FaMTEB: Massive text embedding benchmark in Persian language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11441–11468, Suzhou, China. Association for Computational Linguistics.

Appendix roadmap

Roster and methodology: App. A (legacy IT/JA/HI), App. B (three-tier composition), App. C (task taxonomy), App. F (paper-roster specifications). *Aggregation alternatives*: App. D (multilingual-only per-language), App. E (language-macro LOO), App. K (alternative aggregators incl. Borda). *Same-pipeline replication*

and pipeline-mix audit: App. Q (Swahili-Belebele own-run), App. R (`nemotron-8b` alt-chunker), App. T (per-cell public-vs.-own-run agreement). *Downstream and deployment*: App. G (full MTEB), App. H (RAG by query type), App. I (Pareto frontiers), App. N (deployment), App. P (recipe lift), App. S (hybrid RRF), App. V (native-speaker annotation). *Robustness*: App. L (threshold sensitivity), App. M (leave-one-out τ), App. O (per-task variance / saturation control), App. J (pipeline validation pass rates). *Downstream baselines*: App. U (closed-book RAG baseline). *Artifacts*: App. W (reproduction map: result $\rightarrow$ sub-project $\rightarrow$ entry point). *Dedicated benchmarks*: App. X (per-language fork coverage and rank concordance). *Other task categories*: App. Y (classification/STS leaderboard-only audit). *Snapshot robustness*: App. Z (replication on a 2026-07 leaderboard snapshot).

A Legacy Rank-Inversion Analysis (IT/JA/HI Paper-Roster Subset)

For continuity with smaller-roster analysis, Table 6 reports the IT/JA/HI rank correlation breakdown computed on the legacy IT/JA/HI paper-roster subset (n=13–15, distinct from the 7-model restricted tier of §3.1; the legacy subset is the 17-model paper roster intersected with models that submit to the relevant IT/JA/HI MTEB tasks), including Belebele-only single-task controls. The pattern is consistent with the broader audit (Table 2) but at smaller statistical resolution.

Lang	n	τ [95% CI]	p_{adj}	Inversions
Lang-avg over all available MTEB tasks (headline):				
IT	14	0.45 [0.00, 0.83]	0.040	25/91 (27.5%)
JA	15	0.47 [0.00, 0.79]	0.028	28/105 (26.7%)
HI	15	0.60 [0.16, 0.92]	0.004	21/105 (20.0%)
Belebele-only NDCG (uniform single-task control):				
IT‡	14	0.45	n/a	25/91 (27.5%)
JA‡	14	0.43	n/a	26/91 (28.6%)
HI‡	14	0.47	n/a	24/91 (26.4%)
IT†	15	0.28 [−0.22, 0.73]	0.169	38/105 (36.2%)

Table 6: Kendall τ between MTEB rank and language-specific rank on the legacy IT/JA/HI paper-roster subset (10,000 model-resample CIs); p_{adj} uses Benjamini–Hochberg FDR at q=0.05 over the nine-test family actually corrected in the released statistics: the three IT/JA/HI lang-avg Kendall-τ tests together with three mean-ratio and three anchor-pair bootstrap tests. The IT† row sits outside that family; its p is the value computed in the separate with-`nemotron-8b` analysis. † Italian with `nemotron-8b` included: τ halves and is no longer significant. Japanese and Hindi τ remain stable under leave-one-out (Appendix M). ‡ Belebele-only single-task control rows use the same task across IT/JA/HI to defuse the task-coverage asymmetry; computed over the same n=14 model set as §4.2. Single-task τ values cluster within 0.04 of each other (HI 0.47, IT 0.45, JA 0.43), narrower than the headline spread, consistent with task-coverage asymmetry contributing to the headline cross-language τ differences. p_{adj} omitted for Belebele-only rows (confirmatory follow-ups outside the FDR family).

B Three-Tier Roster Composition

For reproducibility we publish the full membership of each tier in `analysis/roster_membership.csv`, which lists every (model, tier, language) row with the model's global MTEB(Multilingual) rank and per-language local rank used in Tables 2–4.

Restricted (n=7)**.** The paper-roster models with MMTEB(Multilingual) Retrieval submissions at the 2026-05 snapshot (§3.1): `Qwen3-0.6B`, `Qwen3-4B`, `Qwen3-8B`, `nemotron-8b`, `harrier-0.6b`, `bge-m3`, `ml-e5-small`. `ml-e5-small` lacks leaderboard results on the tasks covering `ita`, `hin`, `ben`, `deu`, and `fas`, giving per-language n=6 in those languages (Table 2).

Curated (n=25)**.** Restricted plus 18 deployment-relevant additions, listed with global rank in the 25-model curated tier: `KaLM-Embedding-Gemma3-12B` (1), `pplx-embed-v1-4b` (Eslami et al., 2026) (5), `inf-retriever-v1` (7), `F2LLM-v2-14B` (Zhang et al., 2026b) (8), `Seed1.6-embed` (9), `F2LLM-v2-8B` (10), `pplx-embed-v1-0.6b` (11), `granite-311m` (12), `jina-v5-small` (13), `F2LLM-v2-4B` (15), `voyage-3.5` (Voyage AI, 2025) (16), `jina-v5-nano` (17), `zembed-1` (ZeroEntropy, 2026) (18), `embeddinggemma-300m` (Vera et al., 2025) (19), `BOOM_4B_v1` (Zhang et al., 2026a) (20), `BidirLM-Omni-2.5B-Embedding` (Boizard et al., 2026) (21), `GritLM-7B` (Muennighoff et al., 2025) (22), `BidirLM-1B-Embedding` (23). These 18 were selected by hand as deployment-relevant; §3.1 states on what basis, and the frozen 25-model list is released in `analysis/roster_membership.csv`, so

Roster variant	n	Median τ	Median inv%
Headline curated	25	0.34	33.0
No instruction-tuned	20	0.33	33.3
$\geq$10-lang coverage	25	0.34	33.0
Family-balanced	18	0.35	32.7

Table 7: Curated-tier inversion rate under alternate roster definitions. *No instruction-tuned* excludes the five models requiring a query-side instruction prefix at inference (`GritLM-7B`, `KaLM-Embedding-Gemma3-12B`, `nemotron-8b`, `inf-retriever-v1`, `BOOM_4B_v1`), giving $25 - 5 = 20$. *$\geq$10-lang coverage* restricts to models with at least 10 of the 17 audit languages covered (no models drop at this threshold). *Family-balanced* keeps one model per backbone family (largest retained for multi-model families: 7 models excluded, 18 retained). All four variants give median inversion in $[32.7\%, 33.3\%]$, ruling out cluster cherry-picking, instruction- tuning composition, and family duplication as drivers of the headline.

Subset	n	Definition
Paper roster	17	all evaluated models (App. F)
Legacy IT/JA/HI	13–15	paper roster $\cap$ models submitting to the relevant IT/JA/HI MTEB tasks (App. A)
Deployment	15+2	paper roster $\cap$ 24 GB-fitting models, plus `jina-v3`, `pplx-embed-v1-0.6b` (§3.3)
Task-probe (Belebele/MIRACL)	14	paper roster minus `nemotron-8b` (§4.2)
Task-probe (Wikipedia)	12	above minus 2 models lacking Wikipedia scores (§4.2)
Swahili–Belebele	15	paper roster minus 2 English-only baselines (App. Q)

Table 8: Named model subsets used outside the three-tier system of Table 1, with size and defining rule. The deployment and Swahili–Belebele subsets share n=15 but are defined differently.

the tier is defined by that released list rather than by a mechanical rule.

Extended (n=28–57, language-dependent). All publicly indexed submissions to the 18 MMTEB(Multilingual) retrieval tasks at the 2026-05 snapshot, intersected with each language's available evaluations. `granite-97m-r2` (extended global rank 28) is the cluster member that appears only in the extended tier; per-language inclusion flags are released in `analysis/extended_roster.csv`.

Roster sensitivity. The headline curated-tier inversion rate is robust to alternate roster definitions (Table 7): excluding instruction-tuned models, restricting to models with ≥ 10 language coverage, and one-model-per-family balancing all reproduce the headline within 0.7 percentage points.

C Task Taxonomy for the Hidden-Failure Profile

The 18 retrieval tasks aggregated by MTEB(Multilingual) Mean(Task, Retrieval) classify as follows. *English-derived* (11 tasks; corpus is English or English-language transformations): `ArguAna` (Wachsmuth et al., 2018; Thakur et al., 2021), `TRECCOVID` (Roberts et al., 2021), `SCIDOCS` (Cohan et al., 2020; Thakur et al., 2021), `StackOverflowQA` (Muennighoff et al., 2023), `HagridRetrieval` (Kamalloo et al., 2023), `LegalBenchCorporateLobbying` (Guha et al., 2023), `AILAStatutes` (Bhattacharya et al., 2020) (Indian legal statutes in English), `LEMBPasskeyRetrieval` (Zhu et al., 2024) (English synthetic long-context), `SpartQA` (Mirzaee et al., 2021), `TempReasonL1` (Tan et al., 2023), `WinoGrande` (Sakaguchi et al., 2021). *Multilingual* (7 tasks). This group is defined by *not* being English-derived rather than by within-task language mixing: five are multilingual-suite tasks whose corpora span languages, and two — `CovidRetrieval` (Chinese) and `TwitterHjerneRetrieval` (Danish) — are monolingual non-English. The members are `BelebeleRetrieval` (Bandarkar et al., 2024) (122 languages), `MIRACLRetrievalHardNegatives` (Zhang et al., 2023) (18 languages), `MLQARetrieval` (Lewis et al., 2020) (7 languages), `WikipediaRetrievalMultilingual` (Enevoldsen et al., 2025), `StatcanDialogueDatasetRetrieval` (Lu et al., 2023) (EN/FR bilingual), `CovidRetrieval` (Long et al., 2022) (Chinese), and `TwitterHjerneRetrieval` (Holm, 2024) (Danish). The full taxonomy with per-task rationale is released in `analysis/task_taxonomy.csv`.

Robustness to a strict 5-task multilingual group. Because two of the seven are monolingual non-English, we re-run the cluster-profile contrast on the strict 5-task group that drops them (`BelebeleRetrieval`,

Model	$\overline{\text{eng}}$	$\overline{\text{multi}}$	Δ
Hidden-failure cluster (§4.4)			
harrier-0.6b	0.711	0.684	+0.026
granite-97m-r2	0.598	0.591	+0.006
granite-311m	0.632	0.658	−0.026
Seed1.6-embed	0.624	0.699	−0.075
inf-retriever-v1	0.621	0.717	−0.096
Cluster median	0.624	0.684	−0.026
Non-cluster median (n=21)	0.588	0.706	−0.120

Table 9: Per-task profile of the hidden-failure cluster vs. the non-cluster curated tier (n=21). $\overline{\text{eng}}$ is the mean NDCG@10 over 11 predominantly-English MMTEB retrieval tasks; $\overline{\text{multi}}$ is the mean over 7 multilingual tasks (taxonomy in App. C). All five cluster models sit in the top 12 by Δ across the 26 evaluated models (curated tier + `granite-97m-r2`); cluster median Δ exceeds non-cluster median by 0.094 NDCG (Mann–Whitney one-sided U=91, p=0.005). Δ is computed from unrounded values and need not equal the difference of the two displayed columns. Scale: NDCG@10 on $[0, 1]$.

`MIRACLRetrievalHardNegatives`, `MLQARetrieval`, `WikipediaRetrievalMultilingual`, `StatcanDialogueDatasetRetrieval`). Cluster median Δ is -0.006 against non-cluster -0.088 (Mann–Whitney one-sided U=97, p=0.0009). The result is not simply "robust": the cluster-vs-non-cluster gap *narrows* from 0.094 to 0.083 while significance *strengthens* (p 0.005→0.0009, U 91→97). On the two dropped tasks the cluster is not materially advantaged (`CovidRetrieval` 0.801 cluster vs. 0.798 non-cluster; `TwitterHjerneRetrieval` 0.668 vs. 0.724), so dropping them sharpens the contrast rather than manufacturing it.

Robustness to a stricter English-derived taxonomy. Three of the eleven English-derived tasks are arguable: `AILAStatutes` is Indian legal text in English, `LEMBPasskeyRetrieval` is synthetic, and the reasoning tasks `SpartQA`/`TempReasonL1`/`WinoGrande` are English text but stylistically distinct from retrieval. We re-run the cluster-profile Mann–Whitney test using a strict 6-task subset (`ArguAna`, `TRECCOVID`, `SCIDOCS`, `StackOverflowQA`, `HagridRetrieval`, `LegalBenchCorporateLobbying`). The result holds: U=83, one-sided p=0.024, with cluster median Δ=+0.083 vs. non-cluster Δ=+0.054. Both groups have positive deltas under the strict taxonomy (most curated-tier models score higher on canonical English retrieval than on multilingual tasks), and the cluster's relative English advantage remains larger under this stricter classification.

Cluster characterisation: not driven by instruction-tuning. A natural confound for the eng-vs-multi gap is that the cluster might simply be the subset of models without retrieval-task instruction prefixes (since instruction-tuned retrievers are typically trained on heterogeneous, multilingual instruction-following data). We tested this directly. The 5-model cluster contains one instruction-tuned model (`inf-retriever-v1`) and four non-instruction-tuned models (`granite-311m`, `granite-97m-r2`, `harrier-0.6b`, `Seed1.6-embed`). Across the 21 evaluated models with disclosed instruction-tuning status (curated tier ∪ `granite-97m-r2`, minus the F2LLM family, `voyage-3.5`, and `zembed-1`, which do not disclose; 5 cluster, 16 non-cluster), Fisher's exact test gives an odds ratio of 1.33 (two-sided p=1.00). The cluster is not enriched for either instruction-tuning category; the eng-vs-multi gap is not a proxy for instruction-tuning status. Full per-model metadata (backbone family, parameter count, embedding dim, training languages where disclosed) is released as `analysis/cluster_metadata_full.csv`.

Cluster characterisation: not driven by size or architecture. We further checked whether parameter scale or backbone family explains the cluster. The five cluster members span 97M to ∼7B parameters (a 72× range) and four distinct model families spanning both encoder and decoder architectures: IBM Granite BERT-style encoders (`granite-311m`, `granite-97m-r2`), a decoder-only transformer with last-token pooling (`harrier-0.6b`), a Bytedance decoder-LLM embedder (`Seed1.6-embed`), and an Infly decoder-LLM retriever (`inf-retriever-v1`). Cluster median parameter count (596M, `harrier-0.6b`) sits within the curated-tier interquartile range (584M–7B). Neither parameter scale nor architecture family is a confound for the eng-vs-multi gap signature.

D Multilingual-Only Aggregator: Per-Language Detail

Table 10 reports per-language inversion rate at curated-tier roster size for three aggregators:

ISO	Lang.	Mean(18)	Mean(7)	Mean(11)
ara	Arabic	27.3	**10.0**	32.3
hin	Hindi	26.8	**10.9**	31.5
vie	Vietnamese	20.7	**12.7**	25.0
ben	Bengali	29.0	**13.0**	33.7
spa	Spanish	25.7	**21.7**	28.0
deu	German	29.3	**22.1**	32.6
fas	Persian	34.8	**22.5**	38.0
rus	Russian	33.0	**23.7**	36.0
zho	Chinese	30.3	**23.7**	32.0
ita	Italian	31.9	**24.6**	35.9
tel	Telugu	37.0	**25.0**	40.7
tha	Thai	34.0	**25.3**	37.7
jpn	Japanese	33.3	**26.7**	37.0
kor	Korean	36.0	**26.7**	38.3
ind	Indonesian	37.0	**31.7**	40.0
fra	French	37.3	**33.3**	39.0
swa	Swahili	35.0	35.0	38.0
Median		33.0	**23.7**	36.0

Table 10: Per-language pairwise inversion rate (%) at curated-tier roster size under three aggregators. Mean(18) is the manuscript headline; Mean(7 multilingual) restricts to the 7 multilingual predictor tasks; Mean(11 English-derived) restricts to the 11 English-derived tasks. Bold = lowest inversion rate per row. Sorted by Mean(7) ascending.

Aggregator	Median τ	Median inv%
Mean(11 English-derived)	0.28	36.0
$1/n_t$ reweighted	0.29	35.7
Mean(18) full [headline]	0.34	33.0
Mean(7 multilingual)	0.53	23.7
Lang-macro multi (LOO)	0.70	15.0
Lang-macro 18-lang (LOO)	0.71	14.5

Table 11: Comprehensive aggregator comparison at curated-tier roster size across all 17 audit languages. The two null-or-worse contrasts (top: English-only, $1/n_t$), the headline arithmetic mean, the mechanism evidence (Mean(7) multilingual), and two language-balanced alternatives that substantially improve alignment (bottom). *LOO* = leave-one-out: the target language is excluded from the macro-average when predicting its own rank. Self-inclusive variants of the language-macro aggregators give 0.73/0.71 τ and 13.7%/14.5% inversion, within 0.9 percentage points (pp) of LOO (a gap between aggregate inversion rates, not the median *per-language* difference reported in Figure 4), confirming the result is not an artifact of the target language contributing to its own predictor. $1/n_t$ reweighting fails despite addressing the most obvious coverage asymmetry, localising the disagreement to English-task content rather than broad-coverage dominance.

Mean(18) full (manuscript headline), Mean(7 multilingual), and Mean(11 English-derived). Across all 17 audit languages, Mean(7 multilingual) reduces or matches the inversion rate (Swahili is the sole tie); Mean(11 English-derived) increases it for every language.

E Language-Macro Aggregators (LOO)

Two language-balanced aggregators compute per-language means first and then macro-average across languages: *lang-macro 18-lang* (macro-averaging over all available languages per model) and *lang-macro multilingual-only* (restricted to the 7 multilingual predictor tasks). Both are reported as leave-one-out: when computing the predictor for language X, X is excluded from the macro-average. We additionally test $1/n_t$ *reweighting*, which weights each task by the inverse of its language coverage (Belebele = 1/122, MIRACL = 1/18, etc.), to isolate whether broad-coverage dominance alone explains the disagreement.

Extended-tier generalisation. The language-balanced LOO aggregator also improves the judgment-free extended tier. Over the 53 extended-tier models covering $\geq$14 of the 17 audit languages (baseline median inversion 15.8%, matching the full-roster 15.4% of Table 2), the balanced aggregate cuts median inversion to **7.2%** (median τ 0.68$\rightarrow$0.86), improving 16 of 17 languages (Spanish ties). On the stricter 28-model subset present in all 17 rosters, whose widely-benchmarked composition already yields a low 7.7% baseline, it still improves to 5.8% (τ 0.85$\rightarrow$0.88). Per-language values are released in `extended_aggregator_summary.csv`. The roster-judgment concern thus does not explain the aggregator's effect: it holds on the tier with no curation at all.

Paired uncertainty. Paired within language, the curated-tier improvement over Mean(18) is unambiguous: median paired change in inversion rate is -16.0 points (95% bootstrap CI $[-20.7, -12.0]$, 10,000 language-level resamples; two-sided Wilcoxon signed-rank $p<0.001$, $n=17$), and median paired change in Kendall τ is $+0.320$ ($[0.240, 0.413]$, $p<0.001$). All 17 of 17 languages improve on both metrics, with no ties or reversals.

F Paper-Roster Model Specifications

Model family citations. Paper-roster models by family: multilingual E5 (`ml-e5-small`/`base`/`large`/`large-instruct`) (Wang et al., 2024b), English E5 baselines (`e5-small-v2`, `e5-large-v2`) (Wang et al., 2022), `bge-m3`

Model	Params	Dim	Family
Small			
granite-107m	107M	768	IBM Granite
ml-e5-small	118M	384	E5
jina-v5-nano	212M	1024	Jina AI
e5-small-v2†	33M	384	E5
Medium (278M to 600M)			
ml-e5-base	278M	768	E5
e5-large-v2†	335M	1024	E5
ml-e5-large	560M	1024	E5
ml-e5-large-instruct	560M	1024	E5
bge-m3	568M	1024	BGE
snowflake	568M	1024	Arctic-Embed
harrier-0.6b	596M	1024	Harrier
Qwen3-0.6B	600M	1024	Qwen3
Large (4B to 8B)			
Qwen3-4B	4B	2560	Qwen3
SFR-Mistral	7B	4096	SFR
e5-mistral-7b	7B	4096	E5
nemotron-8b	7.5B	4096	Nemotron
Qwen3-8B	8B	4096	Qwen3

Table 12: The 17 paper-roster models evaluated on language-specific MMTEB tasks (IT/JA/HI), the multilingual RAG benchmark, and the deployment benchmark. **Provenance:** MTEB-agg and per-task MMTEB scores are from the public leaderboard CSV (2026-05); all RAG, deployment, and downstream results are own runs. † English-only baselines, excluded from rank-inversion analysis. The 7 paper-roster models with MMTEB(Multilingual) Retrieval submissions form the restricted tier (§3.1; App. B); the 18 curated-tier additions are listed in App. B.

(Chen et al., 2024), `jina-v5-nano` (Akram et al., 2026), Qwen3 embedding family (`Qwen3-0.6B/4B/8B`) (Zhang et al., 2025), `SFR-Mistral` (Meng et al., 2024), `e5-mistral-7b` (Wang et al., 2024a), `granite-107m` (Awasthy et al., 2025). Backbone citations for `snowflake` (Arctic-Embed-L-v2.0) appear in Yu et al. (2025).

G Full MTEB Results (All 17 Models)

H RAG Results by Query Type

Exploratory: multi-hop vs. factual size correlation. Spearman ρ between log-parameter count and per-query-type NDCG@10, computed across n=18–19 neural retrievers (union of the curated tier and the deployment-tractable roster, BM25 excluded; five evaluation directories without a parameter record were dropped). Multi-hop ρ exceeds factual ρ on 5 of 6 configurations: ja-law (0.413 vs. 0.064), hi-law (0.270 vs. −0.162), it-law (0.245 vs. 0.046), it-fin (0.032 vs. −0.202), and hi-fin (0.480 vs. 0.425); ja-fin is essentially tied (0.338 vs. 0.352). `it-fin/factual` is the only cell with a clearly negative log-params correlation: bigger retrievers do *worse* on Italian factual queries than smaller ones, a language-level pattern consistent with the per-config MTEB-to-RAG propagation analysis of Table 14. We report this as exploratory.

Statistical detail for §5.1. Table 14 collects the two sets of figures behind the downstream claims of §5.1: panel (a) the per-judge significance of the ja-fin keypoint-coverage inversion, and panel (b) the per-configuration strength of MTEB-to-RAG rank propagation.

I Cost–Performance Pareto Frontiers

J Pipeline Validation Pass Rates

K Alternative Aggregation Strategies

A natural follow-up to the rank-inversion finding is whether alternative leaderboard aggregators (beyond MTEB Mean(Task)) would mitigate the disagreement. We compute four score-based alternatives and Borda-count rank aggregation over the 18 MMTEB retrieval tasks, re-run the headline rank-correlation analysis, and report results in Table 16. No alternative score-based aggregator delivers a τ above 0.50 on Japanese or 0.50 on Italian; Median(Task) gives the largest improvement (IT $+0.10\,\tau$, HI $+0.14\,\tau$) and a 10%-trimmed mean gives a smaller but consistent uplift on all three languages, while harmonic mean and min-score actively worsen alignment. The user-side per-language filter remains the simplest dominant fix; we do not propose a new aggregator.

Borda count replication. Beyond score-based aggregators (Table 16), MMTEB also exposes a Borda-count rank aggregator (Enevoldsen et al., 2025). We re-run the headline rank-correlation analysis with Borda rank substituted for Mean(Task) rank, using the same 18 retrieval tasks. Borda gives Kendall τ of 0.38 / 0.18 / 0.31 for IT/JA/HI (95% CI $[-0.01, 0.73]$ / $[-0.25, 0.56]$ / $[-0.10, 0.66]$) with 30.8% / 41.0% / 34.3% inversions, versus 0.45 / 0.47 / 0.60 and 27.5% / 26.7% / 20.0% under Mean(Task). The phenomenon is therefore not specific to arithmetic averaging: under Borda the inversion rate is in fact *higher* for all three languages, substantially so for JA (26.7 to 41.0%) and HI (20.0 to 34.3%).

Model	Pars	MTEB	Italian				Japanese				Hindi					
			Bel.	Wiki	Avg	Δ	Bel.	MRCL	Avg	Δ	Bel.	MRCL	Wiki	MLQA	Avg	Δ
granite-107m	107M	65.0	87.0	85.9	86.5	+21.5	88.1	60.9	74.5	+9.5	73.1	44.3	78.7	51.1	61.8	−3.2
ml-e5-small	118M	77.0	92.3	89.2	90.8	+13.8	91.8	63.9	77.8	+0.8	86.1	55.9	84.4	64.8	72.8	−4.2
jina-v5-nano	212M	72.0	93.8	92.2	93.0	+21.0	93.1	59.1	76.1	+4.1	89.7	53.6	87.6	66.4	74.3	+2.3
e5-small-v2†	33M	41.6	74.2	77.8	76.0	+34.4	40.8	4.8	22.8	n/a	9.6	0.1	10.9	4.0	6.2	n/a
ml-e5-base	278M	79.7	94.2	89.9	92.1	+12.4	92.5	65.1	78.8	−0.9	88.9	59.3	84.2	66.6	74.7	−5.0
ml-e5-large	560M	84.2	95.1	92.4	93.8	+9.6	94.6	70.7	82.7	−1.5	90.7	62.7	87.1	70.5	77.8	−6.4
ml-e5-large-instruct	560M	83.8	92.6	91.7	92.1	+8.3	92.0	66.1	79.0	−4.8	89.0	58.3	87.2	69.2	75.9	−7.9
bge-m3	568M	84.0	93.7	90.1	91.9	+7.9	93.7	73.0	83.4	−0.6	89.3	60.1	87.4	67.5	76.1	−7.9
snowflake	568M	82.3	66.2	64.3	65.3	−17.0	52.4	2.2	27.3*	−55.0	57.8	14.6	56.7	31.4	40.1*	−42.2
e5-large-v2†	335M	56.4	80.7	83.5	82.1	+25.7	49.6	5.1	27.3	n/a	17.3	2.4	21.2	7.4	12.1	n/a
harrier-0.6b	596M	78.0	91.8	90.1	90.9	+12.9	91.0	36.8	63.9	−14.1	86.2	45.7	86.9	64.4	70.8	−7.2
Qwen3-0.6B	600M	77.9	92.4	89.8	91.1	+13.2	90.1	58.6	74.4	−3.5	82.5	45.2	83.4	57.0	67.0	−10.9
Qwen3-4B§	4B	86.2	94.6	93.2	93.9	+7.7	94.1	72.5	83.3	−2.9	91.3	64.2	88.5	82.6	81.7	−4.5
SFR-Mistral‡§	7B	80.3	94.0	n/a	94.0‡	+13.7	90.5	58.0	74.3	−6.0	79.8	48.5	84.4	64.3	69.3	−11.0
e5-mistral-7b‡§	7B	79.6	89.5	n/a	89.5‡	+9.9	62.0	57.8	59.9	−19.7	78.3	47.7	84.5	61.7	68.1	−11.5
nemotron-8b§	7.5B	87.0	28.8	46.3	37.6*	−49.4	95.2	78.0	86.6	−0.4	90.8	67.5	90.6	77.0	81.5	−5.5
Qwen3-8B§	8B	90.4	98.5	93.5	96.0	+5.6	97.9	74.4	86.2	−4.2	96.6	64.8	88.9	84.3	83.7	−6.7

Table 13: Complete MTEB per-task results across all 17 models and three languages. Bel. = BelebeleRetrieval; MRCL = MIRACLRetrievalHardNegatives; Wiki = WikipediaRetrievalMultilingual; MLQA = MLQARetrieval. "n/a" indicates the task is not in the language's MTEB task set or was not evaluated for that model. Δ = lang-avg − MTEB-agg. * = case-study hidden failure (§3.1; `nemotron-8b`/IT: hypothesis-only mechanism; `snowflake`/JA & /HI: documented mechanism, see §4.4). † = English-only reference model. ‡ = IT Avg from BelebeleRetrieval only, the Italian Wikipedia cell being unavailable: for `SFR-Mistral` the own run failed on context-length pre-processing, and for `e5-mistral-7b` the Italian cells are leaderboard-sourced (see §). These IT averages are not directly comparable to other models' two-task IT averages. § marks rows containing **leaderboard-sourced** cells: public per-language values from the released snapshot CSVs rather than own runs. These are `nemotron-8b` Japanese and Hindi (its own-run coverage is Italian only, App. R); `SFR-Mistral` Japanese and Hindi; `e5-mistral-7b` Italian and Hindi plus JA-MRCL; and `Qwen3-4B`/`Qwen3-8B` JA-MRCL together with HI-MRCL, HI-Wiki and HI-MLQA. Every other cell is an own run. This table mixes the two sources by design (§3.1); the 17-language headline audit of Table 2 does not, taking both sides from the same public snapshot. The per-language case-study snapshot CSVs (`case_studies/{jpn,hin}/per_task.csv`) accompany the data release rather than the code release. Italian uses Bel + Wiki; Japanese uses Bel + MRCL; Hindi uses Bel + MRCL + Wiki + MLQA. All values are NDCG@10 × 100.

L Hidden-Failure Threshold Sensitivity

M Leave-One-Out Robustness for Kendall τ

N Full Deployment Benchmark

API embedding endpoints ($/1M tokens + client-side latency). We measured single-query embedding latency from a single WSL2 client (n=50 sequential calls per provider) and pulled public list pricing. The API stack is the deployment choice for practitioners who cannot self-host (compliance, infra constraints); these numbers complement the on-GPU table above for that audience. Rate limits noted are the public free/standard tier as of 2026-05.

O Per-Task Variance Decomposition (Saturation Control)

Section 4.2's task-controlled Spearman comparison rests on holding the task constant across languages. A natural alternative is that single-task evaluation collapses score *variance*, mechanically inflating ρ. Table 21 reports the across-model standard deviation of per-task NDCG@10 for each (language, task) cell on our 17-model working set (English-only baselines and known eval-failures excluded).

P Recipe Lift: Top-3 Selection Quality vs. MTEB-Aggregate

We define a hidden-failure filter (BM25 if it saturates, else Belebele-NDCG@10 screen ≥ 60, then smallest Pareto-frontier model): this predicts a different top-3 candidate set than reading the MTEB aggregate leaderboard. Table 22 simulates each strategy on the 17-model roster and reports the mean true-language NDCG@10 of the picked top-3, plus whether any pick is a documented hidden failure.

Quantity	Setting	Value
(a) Cross-vendor judge agreement, ja-fin keypoint coverage		
Coverage ratio	both judges	1.26×
Wilcoxon p	GPT judge	0.023
Wilcoxon p	Claude judge	0.006
Paired queries	each judge	n=50
(b) MTEB-to-RAG propagation, Spearman ρ		
Italian	it-fin	+0.45
Italian	it-law	+0.44
Japanese, Hindi	other configs	+0.60–0.80

Table 14: Statistical detail for §5.1. **(a)** `bge-m3` vs. `snowflake` on ja-fin, scored for gold-keypoint coverage by each of the two judges of §5.1 over the same paired queries; the coverage ratio favours `bge-m3` and is identical under both judges, and each judge's paired Wilcoxon signed-rank test is significant on its own, so the inversion is not a same-family-judge artifact (Zheng et al., 2023). **(b)** Per-configuration MTEB-to-RAG Spearman ρ as reported in §5.1; the two Italian configurations fall below the Japanese and Hindi ones (ja-fin, ja-law, hi-fin, hi-law), so retrieval-step ranking transfers less reliably downstream in Italian.

Lang/Domain	Factual	Multi-hop	Summ.	Unans.
ja/finance	100%	95%	80%	100%
ja/law	99%	99%	97%	38%†
hi/finance	99%	59%	49%	100%
hi/law	98%	94%	89%	20%†
it/finance	100%	82%	86%	100%
it/law	98%	95%	77%	100%

Table 15: Stage-3 to Stage-5 LLM-validation pass rates per (language, domain, query type). Factual / multi-hop / summarisation: fraction of generated queries passing groundedness + reference-coverage validation. Unanswerable: fraction confirmed unanswerable against the corpus. † Low law-unanswerable rates reflect shared legal vocabulary across cases (court names, common charges) which lets the validator find some corpus document plausibly answering an intentionally fictitious query; finance entities are more lexically unique. The released dataset includes only queries that pass all gates. No validation log is released; each rate is recomputed as the ratio of released queries of that type in the configuration's `full_queries.jsonl` to the per-type target in its `config.yaml`.

Q Own-Run Swahili-Belebele Replication

R nemotron-8b Alt-Chunker Sweep on IT-Wiki

S Hybrid Sparse+Dense (RRF) on Hi-Law and Ja-Fin

T Predictor-Outcome Pipeline Alignment

The 17-language headline audit (Table 2) takes both sides from the same public MMTEB snapshot (§3.1). The IT/JA/HI case studies do mix pipelines: they pair a public-leaderboard *predictor* (MTEB-aggregate / MMTEB Mean(Task)) with own-run *outcomes* (per-language NDCG@10). To verify that this cross-pipeline mix does not drive those results, we audit per-task agreement between own-run and public-leaderboard CSVs for the 17 models on every IT/JA/HI MTEB task in the public per-language CSVs (75 (model, task) cells; published alongside the artifact). Together with the same-pipeline replications referenced in §4.4 (the task-controlled probes of §4.2, the Swahili-Belebele audit in App. Q, and the `nemotron-8b` alt-chunker sweep in App. R), the per-cell delta below is *evidence for* the hidden-failure pattern, not a pipeline artifact: the public leaderboard reports systematically higher scores precisely on the cells where own-run replication detects collapse.

U Closed-Book RAG Baseline

To isolate the retrieval contribution from the answerer's parametric memory, we ran every answerable query through GPT-5.4-Mini (`temperature=0`, `max_completion_tokens=400`) with *no* retrieved context, then scored the resulting answer against the gold keypoints under the same strict numeric rubric and the same OpenAI judge (`temperature=0`, `max_completion_tokens=16`) used for the retrieval-augmented evaluations. Total cost was $3.92 across 4,838 API calls; logs and code are released. The closed-book ablation is single-judge by design: its purpose is to lower-bound parametric memory, not to triangulate judges across vendors.

V Independent Native-Speaker Annotation

To validate the LLM-judge ranking with non-LLM evidence, we ran independent native-speaker annotation on a 50-query subset (factual queries only) of the hi-fin and it-fin cells from §3.2. Two native speakers per language (graduate-level proficiency, volunteer participants) scored each (query, retriever) answer pair on the keypoint coverage rubric of §3.2: 2 = fully covered, 1 = partially covered, 0 = not covered. Annotators were blind to retriever identity (answers labelled only "A" / "B", order ran-

Aggregator	IT τ (inv%)	JA τ (inv%)	HI τ (inv%)
Mean(Task) [headline]	+0.362 (31.9)	+0.333 (33.3)	+0.464 (26.8)
Min-score (worst-task)	+0.130 (43.5)	−0.087 (54.3)	+0.087 (45.7)
Median(Task)	**+0.457** (27.2)	+0.327 (33.7)	**+0.601** (19.9)
Harmonic mean	+0.203 (39.9)	−0.007 (50.3)	+0.145 (42.8)
Trimmed mean (10%)	+0.391 (30.4)	**+0.360** (32.0)	+0.493 (25.4)

Table 16: Kendall τ between alternative MTEB-aggregate predictors and the public per-language lang-avg of Table 2, all computed over the 18-task MMTEB(Multilingual) Retrieval subset at curated-tier roster size (n=24 for IT and HI, n=25 for JA). Bold = best per column. Median(Task) is the only score-based aggregator that materially improves alignment (IT +0.10 τ, HI +0.14 τ); trimmed-mean gives a small consistent uplift; min-score and harmonic mean actively worsen alignment. Among aggregators we tested, only the language-macro alternative (App. E, Table 11) reaches τ=0.71 across all 17 audit languages.

θ_{lang}	θ_{MTEB}	Identified set	$\lvert S\rvert$
40	65	nemotron/it; snowflake/ja	2
45	70	nemotron/it; snowflake/ja; snowflake/hi	3
50	**75**	**nemotron/it; snowflake/ja; snowflake/hi**	**3**
55	80	nemotron/it; snowflake/ja; snowflake/hi	3
60	85	nemotron/it	1
Principled rank-based criterion (threshold-free): MTEB-rank top-50% *and* lang-rank bottom-25% (snowflake/it, snowflake/ja, snowflake/hi, nemotron/it)			**4**

Table 17: Case-study hidden-failure set (§3.1) under five $(\theta_{\text{lang}}, \theta_{\text{MTEB}})$ heuristic thresholds (top block) and under the threshold-free rank-based criterion (bottom block). The principled rank-based criterion is a strict superset of the heuristic, additionally identifying `snowflake/it` (NDCG 65.3, ranked 14/15: second-worst on Italian despite being above the 50 absolute cutoff). All four cases are mechanistically consistent with the documented `snowflake` European-language training distribution and the `nemotron-8b` hypothesis-only Italian case. All thresholds and scores are NDCG@10 × 100.

domised per query) and to all LLM-judge scores; per-answer coverage = $\sum$keypoint scores$/(2 \times n_{\text{keypoints}})$. Annotators worked independently and did not discuss queries until both submitted final files; identities are withheld and only aggregated scores are released.

Hindi (large: bge-m3, small: snowflake). Annotator-averaged keypoint coverage: large **0.481**, small **0.247**; Δ_{human}=**0.234**, paired Wilcoxon p=**0.0012**. Direction agrees with both single-vendor LLM judges (GPT Δ=0.211, Claude Δ=0.248).

Italian (large: jina-v5-nano, small: snowflake). Annotator-averaged keypoint coverage: large **0.526**, small **0.527**; $\Delta_{\text{human}}\approx$0, p=**0.877**. The Italian RAG retriever contrast is not robustly distinguishable on this 50-query slice across any judge (human, GPT, Claude).

Drop:	IT	JA	HI
None (paper baseline)	0.451	0.467	0.600
nemotron-8b	n/a[a]	0.407	0.560
snowflake	0.564	0.560	0.714
ml-e5-small (small anchor)	0.436	0.495	0.670
Qwen3-8B (large anchor)	0.359	0.407	0.560

Table 18: Leave-one-out Kendall τ. Italian stays in 0.36–0.56, Japanese in 0.41–0.56, Hindi in 0.56–0.71. Removing `snowflake` *strengthens* τ in every language, consistent with `snowflake` being the single most-disagreement-inducing model. The Italian τ is the most sensitive to dropping the large-anchor model (`Qwen3-8B`), reflecting Italian's small task base. [a] `nemotron-8b` is excluded from the Italian baseline (Table 6); the with-`nemotron-8b` Italian τ is reported in that table's † row.

Inter-annotator agreement. Spearman ρ on per-answer coverage: **0.949** [0.918, 0.974] on Hindi, **0.982** [0.958, 0.996] on Italian (n=100 each).

W Reproducibility and Artifact Reproduction Map

Compute and licences. All evaluation was conducted on a single NVIDIA RTX 3090 (24 GB); total compute was approximately 35 GPU-hours. The per-run timers in the code release account for roughly 8 of those hours; the remainder (MMTEB sweeps, index builds, and model downloads) was not captured by a timed harness. Code (MIT licence), evaluation scripts, model outputs, public-leaderboard CSV snapshots (2026-05), and the Multilingual RAG datasets (CC-BY-4.0) are released at the two URLs given in §1. The code release comprises six sub-projects with 162 Python files and 34 unit-test files, plus 96 released CSV artifacts.

Reproduction map. Table 29 maps the main results to their regeneration entry points; each sub-project README documents run order and requirements.

X Dedicated Per-Language Benchmark Concordance

For the dedicated per-language MTEB derivatives cited in §2, we measure, from the frozen results

		Query time				Index build	
Model	**Params**	**QPS**	**VRAM**	**p50**	**p99**	t_{build}	**VRAM**
granite-107m	107M	3,474	288 MB	5.0	5.5	0.5 s	417 MB
ml-e5-small	118M	2,758	593 MB	8.5	9.6	0.4 s	705 MB
e5-small-v2	33M	2,275	329 MB	7.9	8.8	0.5 s	832 MB
e5-large-v2	335M	498	2,324 MB	16.3	19.1	n/a	n/a
ml-e5-base	278M	2,041	1,442 MB	7.9	8.7	0.6 s	1,645 MB
jina-v5-nano	212M	622	1,363 MB	31.4	35.4	6.0 s	5,609 MB
jina-v3	572M	401	2,205 MB	47.2	52.3	n/a	n/a
ml-e5-large-instruct	560M	627	1,326 MB	13.6	15.3	1.7 s	1,621 MB
ml-e5-large	560M	802	3,042 MB	14.0	15.2	1.7 s	3,267 MB
bge-m3	568M	830	3,073 MB	14.9	18.0	6.4 s	4,697 MB
snowflake	568M	732	3,069 MB	14.0	15.4	6.5 s	4,698 MB
harrier-0.6b	596M	191	2,049 MB	26.4	28.5	18.5 s	3,628 MB
Qwen3-0.6B	600M	159	2,379 MB	28.5	33.2	18.8 s	**11,617 MB**
pplx-embed-v1-0.6b	596M	189	3,705 MB	30.9	34.4	n/a	n/a
Qwen3-4B	4B	40	8,583 MB	27.4	28.7	**OOM***	n/a
SFR-Mistral	7B	19	14,523 MB	44.6	45.1	**OOM***	n/a
e5-mistral-7b	7B	19	14,524 MB	44.1	44.4	**OOM***	n/a
Qwen3-8B	8B	n/a	n/a	n/a	n/a	**OOM***	n/a

Table 19: Full deployment benchmark on a single RTX 3090 (24 GB; CUDA 12.8, PyTorch 2.8). *Query-time*: 3 trials; 1,000-document corpus already encoded; 1,744 queries; latency in ms. *Index build*: wall-time + peak VRAM during one ja-fin corpus encode pass (250 documents, `batch_size=32`, fp16 autocast, same SentenceTransformer used for both passages and queries; absolute build times scale near-linearly with corpus size, but relative cost ranking is stable). * **OOM** indicates the model could not complete an index-build pass on the 24 GB card. Notably Qwen3-0.6B already uses 11.6 GB during build (an $\approx$2.5$\times$ overhead vs. similar-scale 568M models), which leaves insufficient headroom for the 4B–8B tier on commodity GPUs. **Practical takeaway: index-build VRAM, not query-time VRAM, is the binding constraint for self-hosted deployment on commodity hardware.**

API	**Dim**	**$/1M**	**p50**	**p99**
text-embedding-3-small	1536	0.020	254	2256
text-embedding-3-large	3072	0.130	650	1572
cohere-embed-multilingual-v3	1024	0.100	320	363
gemini-embedding-2	3072	0.150	478	587

Table 20: API embedding deployment numbers: list pricing per 1M input tokens and measured client-side latency (ms) over n=50 sequential single-query calls from a single WSL2 client (single region; cold-tier free/standard-tier endpoints). p99 spikes on text-3-small and -large reflect single rate-limit transient retries; production-tier latency is reportedly flatter (we did not measure that tier). Released alongside the artifact.

Lang	**Task**	n	**mean**	**std**	**range**
IT	Belebele	14	86.8	**17.7**	69.7
IT	Wiki	13	85.3	13.4	47.2
JA	Belebele	13	87.2	13.1	45.6
JA	MIRACL	10	55.6	**20.1**	70.8
HI	Belebele	12	85.1	9.9	38.8
HI	MIRACL	10	50.0	13.4	48.1
HI	Wiki	10	82.3	**9.0**	30.9
HI	MLQA	10	60.9	11.3	39.1

Table 21: Per-task across-model standard deviations of NDCG@10. The saturation-only alternative to §4.2 predicts that Belebele should have the smallest std (most saturated), but Italian Belebele std = 17.7 is the *largest* on the IT side; Japanese MIRACL std = 20.1 is the largest overall. Hindi Wiki std = 9.0 is the smallest, so HI-side variance compression is a partial contributor to the IT–HI Wiki ρ=0.937, but the Italian-side variance is comparable to its Belebele-only variance, ruling out full saturation as the explanation. All values are NDCG@10 $\times$ 100.

snapshot and the installed benchmark-registry definitions, (a) how much of the curated roster each covers and (b) how each fork's own ranking relates to the global MTEB-agg ranking vs. our per-language lang-avg ranking (Table 30). Five of the six are present in the registry snapshot; ArabicMTEB is not and is omitted. A model counts as covered with $\geq$50% of the fork's tasks; fork scores average `main_score` within task, then across tasks.

Coverage is sparse: at most 10 of the 25 curated models per fork (18–52 of the 578 models in the whole snapshot), and **no hidden-failure model of §4.4 meets the 50% full-benchmark coverage threshold on any available fork** (all five sit at $\leq$0.25 full coverage). The one partial exception is `inf-retriever-v1`, which covers all *retrieval* tasks of the French and Chinese forks, so it is rankable on the retrieval-only concordance but not on the full-benchmark one. The models for which the aggregate is most misleading are exactly those the dedicated benchmarks do not rank, so fork consultation cannot replace a reliable default surface. Where overlap permits comparison ($n\geq$6), fork rankings track our per-language rankings closely (C-MTEB full-benchmark τ=1.00 against the Chinese lang-avg; ruMTEB retrieval-only τ=0.82 against the Russian lang-avg, exceeding its 0.69 agreement with the global aggregate); MTEB-French correlates weakly with both references on its 8 covered models. Per-model coverage and all pairwise correlations are released in `fork_concordance.csv`.

Y Cross-Category Generalisation (Leaderboard-Only)

We replicate the per-language vs. global rank-agreement audit for the **classification** and **STS** categories of MMTEB(Multilingual), using public leaderboard results only (task `main_score`; no own runs) over the same frozen snapshot. Per-

Lang	Strategy	Mean	HF
IT	Oracle (true top-3)	**94.6**	n/a
IT	MTEB-aggregate top-3	75.8	**Yes**
IT	Belebele-only top-3 (proxy filter)	94.6	no
IT	Recipe (Bel≥60 + Pareto)	94.6	no
JA	Oracle	**85.5**	n/a
JA	MTEB-aggregate top-3	85.5	no
JA	Belebele-only top-3	85.3	no
JA	Recipe	85.3	no
HI	Oracle	**83.1**	n/a
HI	MTEB-aggregate top-3	83.1	no
HI	Belebele-only top-3	83.1	no
HI	Recipe	83.1	no

Table 22: Top-3-pick lift by selection strategy. *Mean* is the mean own-run language-specific NDCG@10 of the picked top-3 (higher is better). *HF* marks whether any picked model is a documented case-study hidden failure (§3.1). The recipe lift is concentrated on Italian: MTEB-aggregate picks `nemotron-8b` (a hidden failure on IT), losing 18.8 NDCG points relative to the oracle (94.6 − 75.8); the recipe avoids the hidden failure and recovers the oracle within rounding. On Japanese and Hindi, all strategies converge on substantially similar top-3 sets. The recipe's practitioner value is therefore primarily as a *hidden-failure filter*, not as a top-pick reranker, in our 17-model sample. All values are NDCG@10 × 100.

Model	MTEB	SWA-Bel	Δ
Qwen3-8B	90.4	86.3	−4.1
nemotron-8b*	87.0	23.8**	−63.2
Qwen3-4B	86.2	78.9	−7.3
ml-e5-large	84.2	88.9	+4.7
bge-m3	84.0	89.9	+5.9
ml-e5-large-instruct	83.8	89.3	+5.5
snowflake†	82.3	26.9	−55.4
SFR-Mistral	80.3	75.1	−5.2
ml-e5-base	79.7	87.2	+7.5
e5-mistral-7b	79.6	54.7	−24.9
harrier-0.6b	78.0	80.5	+2.5
Qwen3-0.6B	77.9	64.4	−13.5
ml-e5-small	77.0	82.9	+5.9
jina-v5-nano	72.0	71.4	−0.6
granite-107m	65.0	65.7	+0.7

Table 23: Own-run Swahili-Belebele NDCG@10 vs. MTEB-aggregate (n=15 working-set models = 17 paper-roster minus the 2 English-only baselines, `e5-small-v2` and `e5-large-v2`). * `nemotron-8b` OOM'd at default batch and retried at batch 16, producing an anomalously low score; we report it for transparency but exclude it from the headline τ. ** artifact value, see text. † `snowflake` on Swahili (rank 7 to 14) satisfies the rank-based case-study hidden-failure criterion (§3.1), extending the documented `snowflake` non-European-language pattern to a 4th language. τ vs. MTEB-agg: 0.30 (excl. `nemotron-8b`, 35.2% inversion) / 0.14 (incl., 42.9%); both more pessimistic than the leaderboard-derived Swahili τ in Table 2. All values are NDCG@10 × 100.

language scores average a task's subsets for the language, then across the category's tasks exposing it ($n_t \geq 2$ required: 11 audit languages qualify for classification, 8 for STS; **reranking exposes <2 tasks for every audit language and cannot be audited at all**, itself an instance of coverage asymmetry). Rosters: all models with ≥80% category coverage ("extended", judgment-free; results stable for thresholds 0.7–1.0), and its intersection with the curated 25 (n=22).

Rank inversions are not retrieval-specific (Table 31): median inversion is 15.6% (classification) and 19.1% (STS) on the curated intersection, with per-language τ FDR-significant in *every* qualifying language, and 10.2% on the broad extended rosters, mirroring the retrieval pattern in which multilingual-first broad submitters show higher agreement (§4.1). The coverage–inversion association appears in classification (extended $\rho = -0.59$, $p = 0.055$) but not STS ($\rho \approx -0.06$), so we treat the mechanism evidence outside retrieval as suggestive. A sanity anchor run through the same code path reproduces the retrieval extended-tier median inversion (15.4%) exactly. Per-language values: `summary_{classification,sts}_*.csv`.

Z Leaderboard Snapshot Robustness

Re-extracting the curated tier from a fresh public results-repository snapshot dated 2026-07-09 (two months after the frozen audit snapshot), holding the roster and the 18-task global definition fixed, reproduces every headline number: median inversion 33.0% (frozen: 33.0%), median τ 0.34 (0.34), coverage correlation $\rho = -0.58$, $p = 0.014$ (frozen -0.59), and the identical hidden-failure cluster (`granite-311m` 16, `harrier-0.6b` 10, `Seed1.6-embed` 6, `inf-retriever-v1` 3; the same 35 curated cells). Six of 17 languages are bit-identical; the largest per-language shift is 0.022 in τ (Persian). All 25 curated models retain complete 18-task coverage. Per-language values, per-model revision hashes, and the extraction script are released (`fresh_snapshot_summary.csv`).

max_seq_length	IT-Wiki NDCG@10
128	49.7
256	46.3
512	46.3
1024	46.3

Table 24: `nemotron-8b` on Italian-Wikipedia under four `max_seq_length` settings (n=1 model, 4 runs). All four variants give NDCG@10 in $[46.3, 49.7]$, well below the 50/75 hidden-failure threshold, confirming the model-failure interpretation. The IT-Wiki collapse is robust to chunking and truncation, ruling out the preprocessing-artifact alternative. The training-data domain-coverage hypothesis (§3.1) remains the residual unconfirmed cause.

Config	Query type	nem-8b	ml-e5-sm	ml-e5-l	Qwen3-4B	gap
it-fin	factual	46.4	67.7	64.1	72.0	−21.3
it-fin	multi-hop	26.3	56.9	61.7	79.4	−30.5
it-fin	summarisation	68.8	78.0	79.8	83.3	−9.2
it-law	factual	18.3	37.6	37.9	47.0	−19.3
it-law	multi-hop	43.1	45.6	44.7	70.6	−2.6
it-law	summarisation	32.7	48.0	61.0	60.3	−15.2

Table 25: NDCG@10 ($\times$100) on Italian RAG configurations by query type (factual / multi-hop / summarisation only); the Table 5 headline differs because it averages over *all* queries (including unanswerable) weighted by query count. *Gap* = `nemotron-8b` − `ml-e5-small` (negative means `nemotron-8b` worse). Averaged across both domains, the gap is −20.3 pp on factual, −16.6 pp on multi-hop, −12.2 pp on summarisation; factual is the largest, partially consistent with the hypothesised encyclopedic domain-coverage gap (§3.1). Per-domain, it-fin multi-hop is the single worst at −30.5 pp.

Config	Model	BM25	Dense	Hybrid
hi-law (BM25-dominated; BM25 leads dense by 16+ pts):				
hi-law	jina-v5-nano	79.0	59.0	**78.0**
hi-law	bge-m3	79.0	61.0	**78.0**
hi-law	Qwen3-8B	79.0	63.0	**78.0**
ja-fin (dense-tractable control; BM25 well behind):				
ja-fin	jina-v5-nano	50.0	**96.0**	80.0
ja-fin	bge-m3	50.0	**93.0**	80.0
ja-fin	Qwen3-8B	50.0	**97.0**	80.0

Table 26: Reciprocal Rank Fusion (Cormack et al., 2009) (k=60) hybrid sparse+dense NDCG@10. On the BM25-dominated hi-law config, RRF recovers $\geq$98% of BM25 with all three dense models. On the dense-tractable ja-fin control, RRF *drops* dense by 13–17 NDCG points (Qwen3-8B 97 to 80). Scale: NDCG@10 $\times$ 100, rounded to whole points; superseding $[0, 1]$, displayed as whole points. **Practitioner verdict: use RRF only when BM25 dominates dense by $\geq$10 NDCG points; not safe as a universal default.**

Slice	n	med. $\lvert\Delta\rvert$	mean $\lvert\Delta\rvert$	$\lvert\Delta\rvert < 5$
Non-HF cells	67	2.8	6.0	58%
HF cells (snowflake, nemotron)	8	40.3	41.3	0%

Table 27: Per-cell absolute disagreement (NDCG points) between own-run and public-leaderboard scores. HF here means case-study hidden failure (§3.1). **Non-HF cells**: median $|\Delta|$ 2.8 NDCG; non-zero residuals reflect a Belebele-subset choice (monolingual `hin_Deva-hin_Deva` in our pipeline vs. the multi-pair average sometimes reported on the public leaderboard) and minor pipeline-version drift, not noise. **HF cells**: large gaps ($|\Delta|$ from 18 to 67 NDCG): this gap *is* the hidden-failure finding, since leaderboard per-task scores for `snowflake/nemotron-8b` are substantially higher than our own-run replication on the same task, confirming that the leaderboard masks the language-specific collapse. The same-pipeline counterparts (§4.2 task-controlled Spearman ρ, the own-run Swahili replication, and the `nemotron-8b` alt-chunker robustness check) all show the same inversion / hidden-failure pattern using own-run on both sides, so the headline result is not driven by the pipeline mix.

Config	n_q	Closed-book	BM25	Best dense	Δ vs. BM25	Δ vs. best dense
it-fin	404	0.145	0.787	0.783 (jina-v5-nano)	−0.642	−0.638
it-law	416	0.194	0.629	0.590 (Qwen3-4B)	−0.435	−0.396
ja-fin	422	0.167	0.503	0.969 (Qwen3-8B)	−0.336	−0.803
ja-law	444	0.162	0.310	0.286 (Qwen3-4B)	−0.148	−0.124
hi-fin	325	0.268	0.923	0.813 (Qwen3-4B)	−0.655	−0.545
hi-law	414	0.358	0.790	0.628 (Qwen3-8B)	−0.432	−0.270

Table 28: Closed-book GPT-5.4-Mini keypoint coverage with no retrieved context, vs. BM25 and best-dense NDCG@10 on the same queries. *Closed-book* is the fraction of gold keypoints the OpenAI judge marked as covered; BM25 / best dense are NDCG@10 of the retrieval step. The two metrics are not on the same scale (coverage scores a post-answer signal; NDCG@10 scores a pre-answer retrieval ranking), so the Δ columns are reference gaps rather than head-to-head comparisons. The qualitative read is robust (all values on $[0, 1]$): closed-book underperforms retrieval on every config; the gap is smallest on ja-law (where retrieval itself is weak, so closed-book and retrieval are both low) and largest on ja-fin (where the best dense retriever is strongest, so the relative cost of skipping retrieval is highest). The Italian retrieval anomaly is discussed in §5.1 and App. H.

Result	Entry point
Three-tier audit (Table 2)	`mteb-ranking-audit/code/` scripts `01–05`
Roster sensitivity (Table 7)	`extended-experiments/exp8_roster_sensitivity.py`
Robustness appendices	`extended-experiments/run_all.py`
Statistics and figures	`analysis/analysis.py`
Per-language MTEB sweeps (Table 13)	`mteb-language-gap/scripts/eval_lb.py`
RAG retrieval (Table 5)	`rag-dataset/scripts/eval_rag.py`, `eval_bm25.py`, `eval_api_embed.py`
Downstream RAG and judges (§5.1)	`rag-dataset/scripts/eval_downstream_rag.py`, `eval_R4b_judge.py`
Closed-book baseline (App. U)	`extended-experiments/exp11_closed_book_baseline.py`
Deployment benchmark (App. N)	`rag-deployment-benchmark/benchmark.py`

Table 29: Reproduction map from released code (six sub-projects) to the paper's results. The RAG datasets (all six configurations) are released separately under CC-BY-4.0 at the HuggingFace link in the Reproducibility statement.

Fork	Registry name	/25	n	τ_{glob}	τ_{lang}
French (fra)	MTEB(fra, v1)	8	8	0.36 / –	0.14 / –
Persian (fas)	MTEB(fas, v2)	6	5	underpowered	
Russian (rus)	MTEB(rus, v1.1)	10	10	0.78 / 0.69	0.56 / **0.82**
Chinese (zho)	MTEB(cmn, v1)	7	7–8	0.90 / 0.93	**1.00** / 0.86
Vietnamese (vie)	VN-MTEB (vie, v1)	5	5	underpowered	
Arabic (ara)	not in registry	–	–	–	–

Table 30: Dedicated-benchmark concordance. **/25**: curated models with $\geq$50% fork-task coverage. τ_{glob}/τ_{lang}: Kendall τ of the fork's ranking vs. the global aggregate and vs. our per-language lang-avg, as full-benchmark / retrieval-only pairs; pairs with $n<6$ are marked underpowered and reported only in the released CSV. No hidden-failure model (§4.4) meets the 50% full-benchmark threshold on any fork; `inf-retriever-v1` is the one partial exception, at full retrieval-task coverage on the French and Chinese forks.

Category	Roster	n_{mod}	n_{lang}	med. inv%	med. τ	FDR
Classif.	curated $\cap$	22	11	15.6	0.69	11/11
Classif.	extended	54	11	10.2	0.80	11/11
STS	curated $\cap$	22	8	19.1	0.62	8/8
STS	extended	53	8	10.2	0.80	8/8
Rerank.	–	–	0	$n_t<2$ everywhere		

Table 31: Leaderboard-only cross-category audit. curated $\cap$: intersection of the curated 25 with models meeting the category coverage threshold. FDR: languages rejecting H_0: τ=0 at q=0.05 after BH correction within (category, roster).

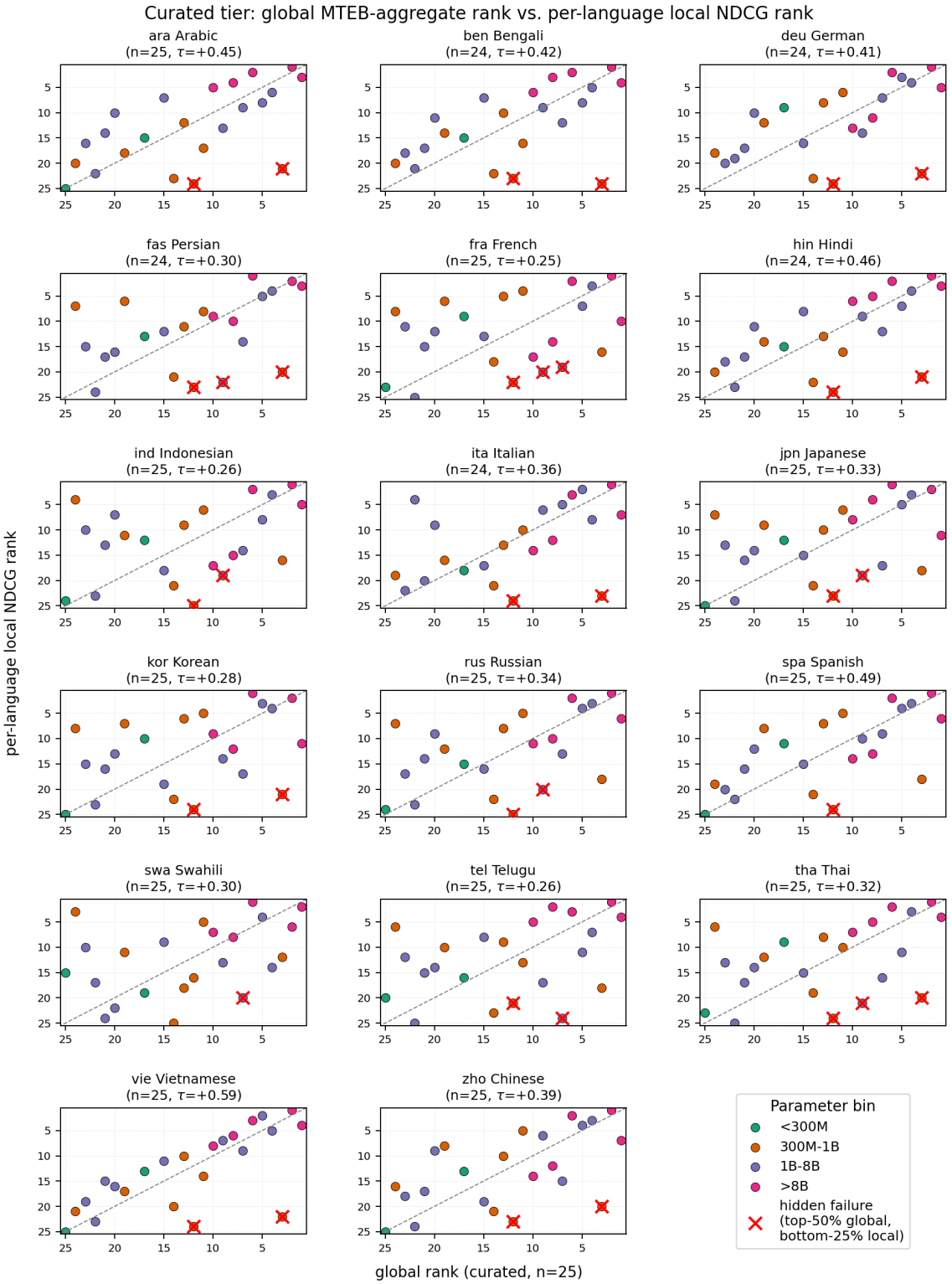


Figure 2: Per-model MTEB global rank vs. language-specific local rank across all 17 audit languages at curated-tier roster size (n=25; the headline tier of Table 2). Points above the y=x diagonal are undervalued by MTEB; points below are overvalued. Red $\times$ marks hidden failures (top-50% global, bottom-25% local; §4.4); 35 hidden-failure cells in total, dominated by `granite-311m` in 16 of 17 languages. Eleven of the seventeen languages have more than half of models *below* the diagonal, i.e. overvalued by MTEB (highest: Vietnamese 60%, Bengali and Italian 58%); only Swahili (56%) and Indonesian (52%) have more than half above. Per-language τ values reproduce Table 2 exactly. Colour: parameter size tier.

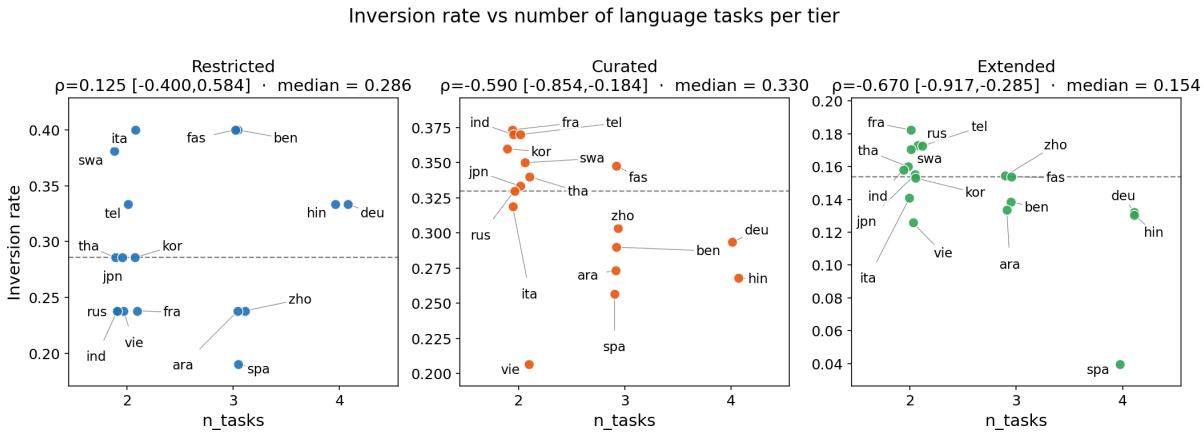


Figure 3: Inversion rate vs. number of MTEB retrieval tasks per language, by roster tier. Each point is one of 17 audit languages. Spearman ρ with 95% bootstrap CI shown per panel. In the curated and extended tiers, languages with fewer tasks show significantly higher inversion rates, the structural prediction of the asymmetric-coverage mechanism (§4.3). Dashed line: median inversion rate within tier.

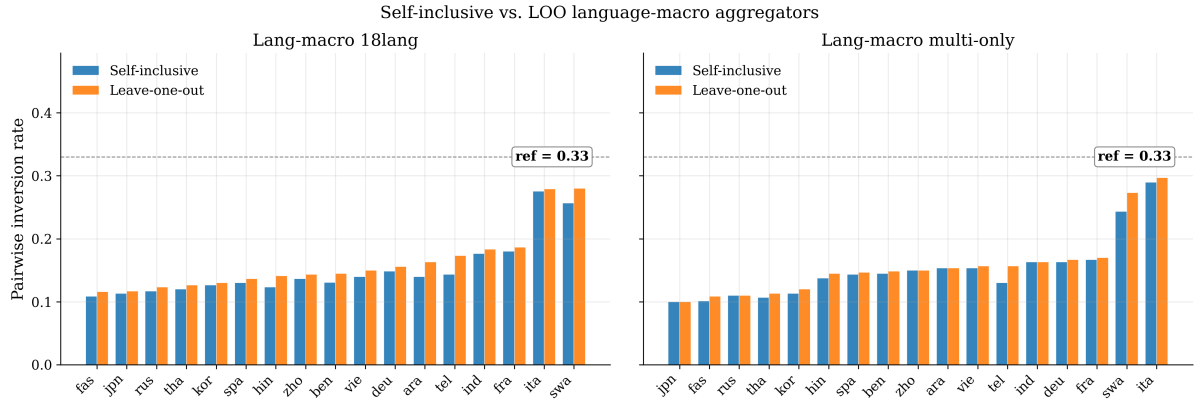


Figure 4: Per-language pairwise inversion rate, self-inclusive vs. leave-one-out, for the two language-macro aggregators. Self-inclusive includes the target language with weight $1/17$ or $1/18$; LOO excludes it. Bars are visually near-identical per-language (median self-inclusive vs. LOO difference ≤ 0.8pp), confirming that the τ=0.71 result is not driven by the target language contributing to its own predictor. Dashed line: the Mean(18) headline inversion rate (33%) for reference.

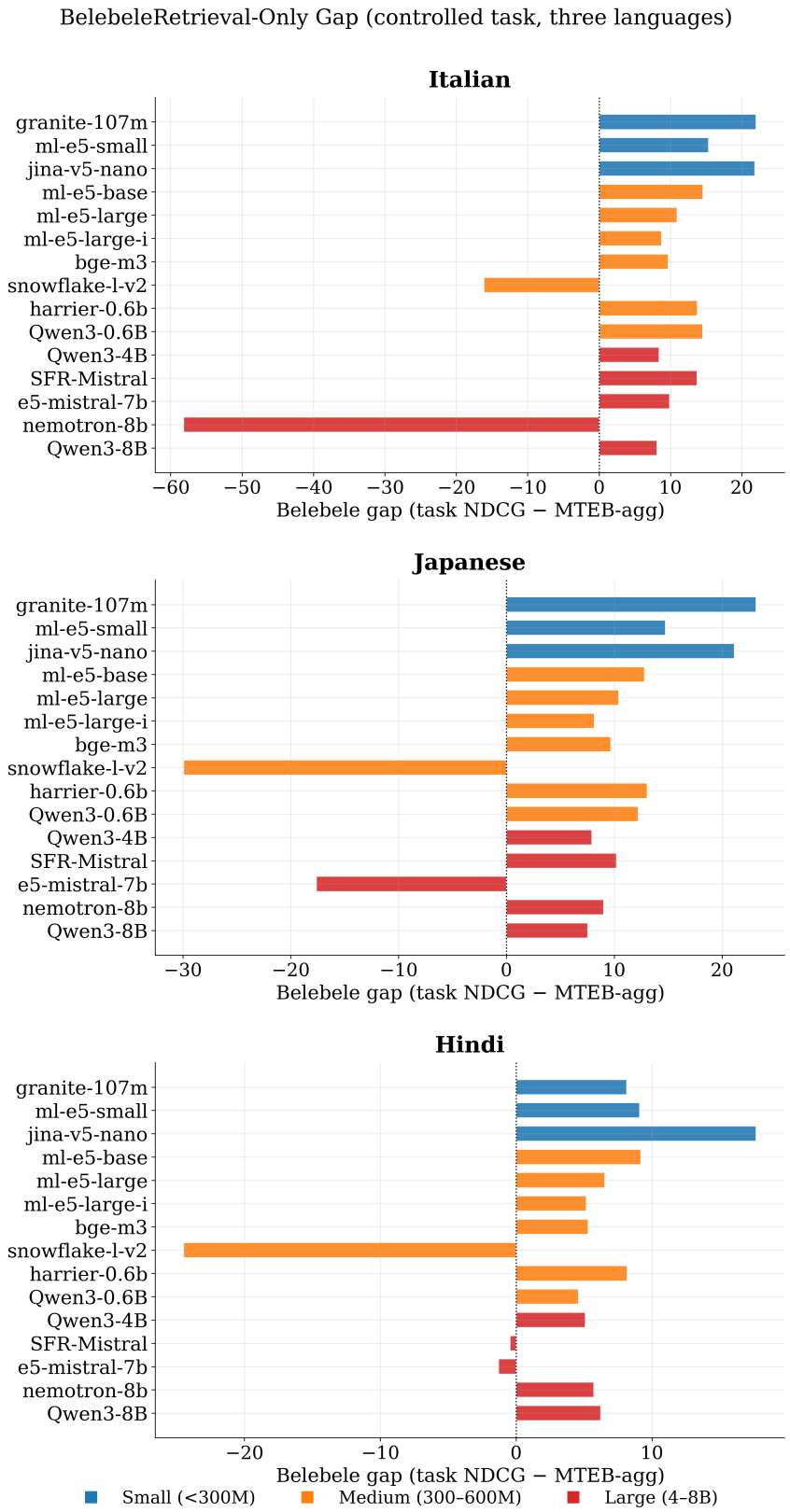


Figure 5: Per-model BelebeleRetrieval-only gap (task NDCG − MTEB-agg) for all 15 evaluated models (the 17 paper-roster models minus the two English-only baselines) in each of the three languages. Holding the task constant collapses cross-language behaviour to a near-uniform pattern: the same models are consistently undervalued (positive gap) and the same models are consistently overvalued (negative gap, including the `snowflake` outlier on non-Latin scripts; the figure labels `snowflake` as `snowflake-l-v2`, a shorthand for the Hugging Face repository `Snowflake/snowflake-arctic-embed-l-v2.0`). Colour denotes parameter tier: small (blue), medium (orange), large (red).

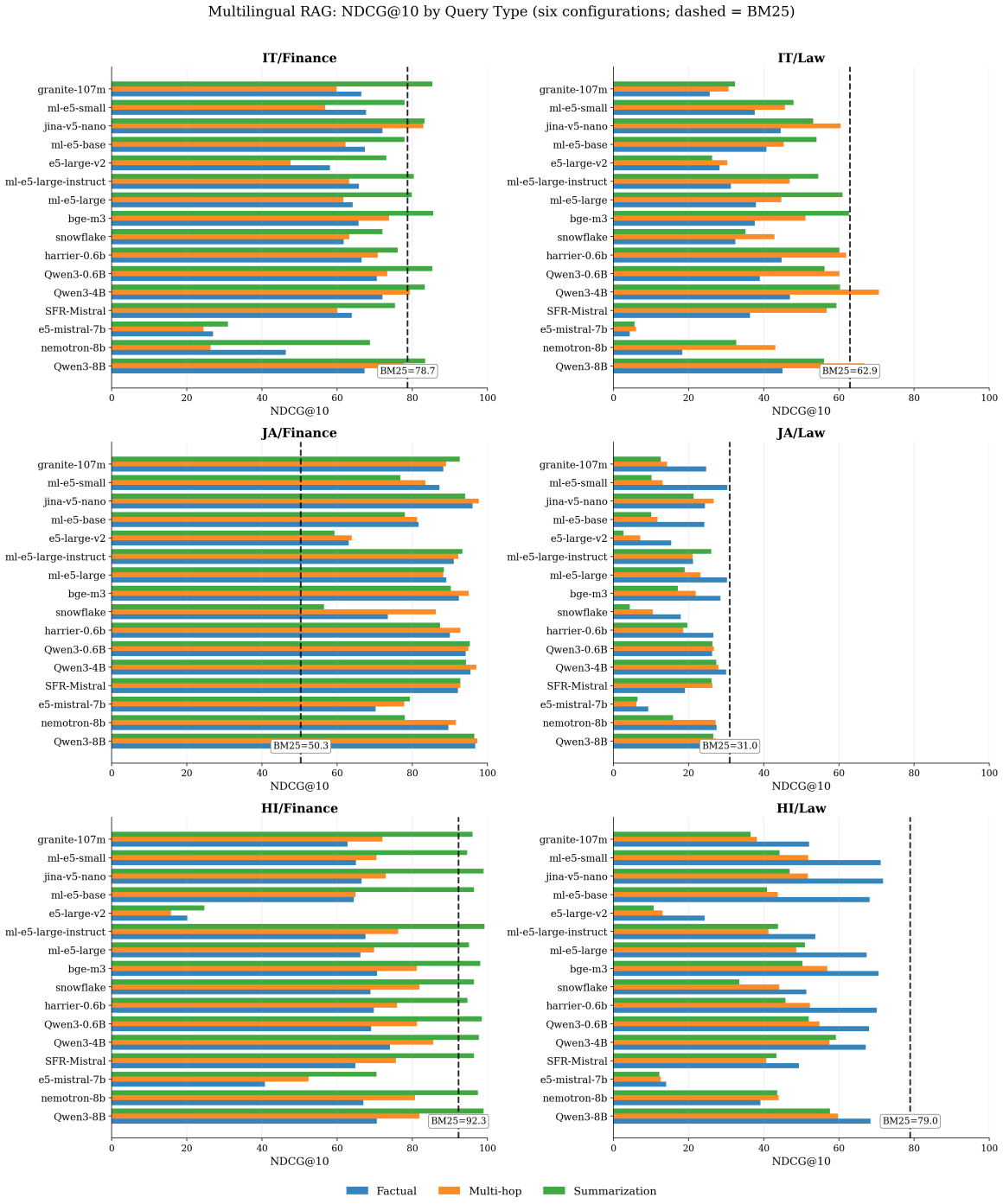


Figure 6: RAG NDCG@10 broken down by query type across all six configurations. Dashed lines show the BM25 baseline per panel. *it-fin*: BM25 (78.7) marginally leads every neural retriever, the best being `jina-v5-nano` at 78.3, a 0.4-point BM25 lead; `nemotron-8b` collapses to 44.0, reproducing the MTEB Italian hidden failure of §3.1. *it-law*: BM25 (62.9) leads all neural retrievers (best neural: Qwen3-4B at 59.0). *ja-fin*: every neural retriever exceeds BM25 (50.3) apart from the `gemini-embedding-2` endpoint (28.6); small models (`ml-e5-small`, `jina-v5-nano`) match large models within 1–7 NDCG points. *ja-law*: every neural model falls below BM25 (31.0), reflecting BM25's lexical-matching advantage on technical Japanese legal text. *hi-fin*: BM25 (92.3) dominates, consistent with its lexical-matching advantage on space-delimited entity names (§3.2). *hi-law*: dense retrieval recovers and small models capture 87–94% of the best large model's NDCG@10.

Cost–Performance Pareto Frontiers (lang-specific in colour, MTEB-agg in grey; bold = on Pareto frontier)

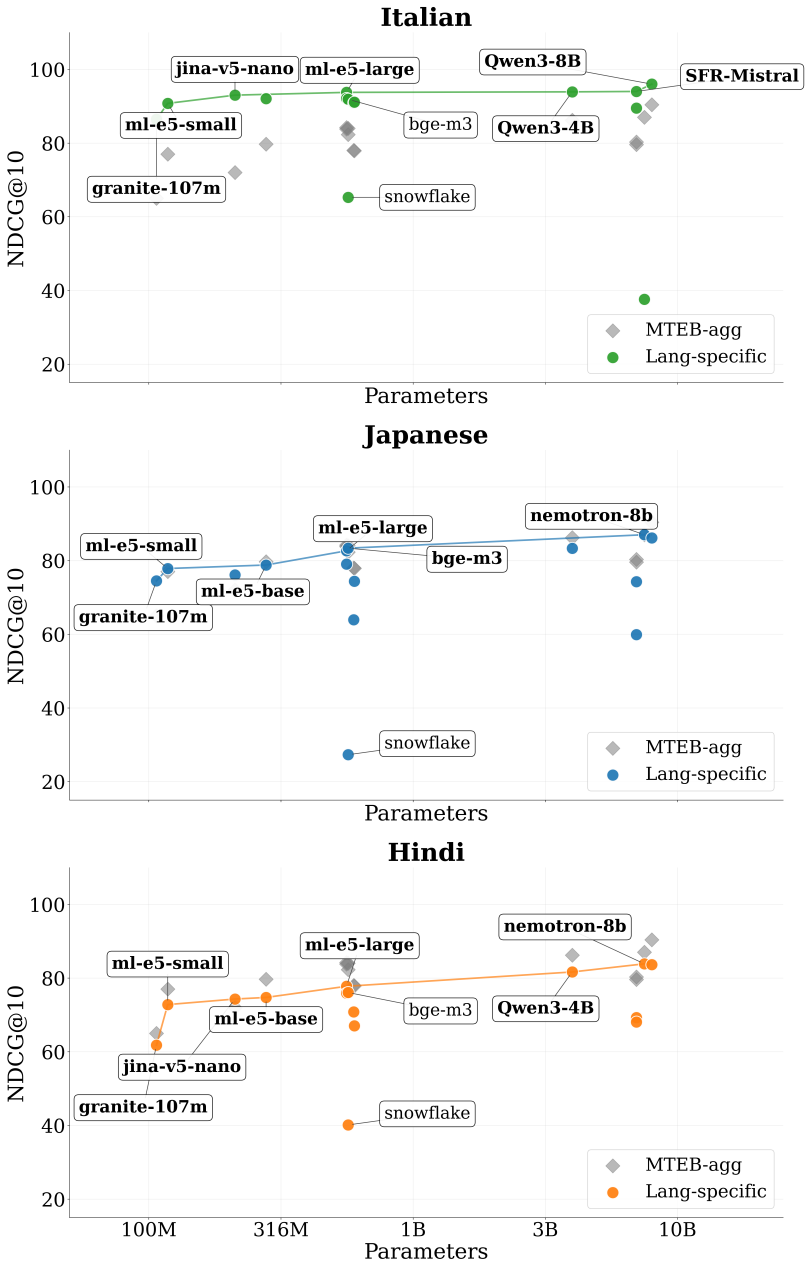


Figure 7: Cost–performance Pareto frontiers for Italian (top), Japanese (middle), and Hindi (bottom). Each point is one of the 17 evaluated models; x-axis is parameter count (log scale), y-axis is NDCG@10. Coloured circles show *language-specific* performance and form the language-Pareto frontier; grey diamonds show the same models' *MTEB-aggregate* score. Models with a coloured circle above a grey diamond at the same parameter count sit on the language-Pareto frontier but not on the MTEB-aggregate frontier: exactly the models a practitioner would miss by selecting on aggregated leaderboard scores. Small models (`ml-e5-small`, `granite-107m`) appear on every language-Pareto frontier; `bge-m3` is a Pareto-efficient mid-tier choice for Japanese only; on Italian and Hindi it is dominated by `ml-e5-large`, which has comparable parameter count and higher language NDCG@10.