

Benchmark Rank Is Not Adaptation Rank: Evaluating Multilingual Sentence Encoders Inside the Data-Selection Loop

Divyajot Singh

Birla Institute of Technology and Science, Pilani
divyajot.5005@gmail.com

Abstract

Multilingual sentence encoders are selected almost entirely by their aggregate rank on massive multilingual benchmarks, and their Matryoshka prefixes are validated with static linear probes. Both practices assume the evaluation data is fixed. In low-budget adaptation the data is not fixed: the encoder’s own geometry and probabilities decide which examples get annotated, and those annotations become the next training set. We audit three massively multilingual encoders—Harrier-OSS-v1-0.6B, Qwen3-Embedding-0.6B, and EmbeddingGemma-300M, spanning 94–100+ languages—inside this loop, crossing three representation treatments, five acquisition policies, five paired seeds, and budgets of 50–500 labels (975 configurations) on two humanitarian text classification testbeds. Multilingual benchmark rank does not survive contact with the loop. The lowest-ranked encoder (EmbeddingGemma, 61.15 MTEB multilingual) attains the best held-out macro-F1 and worst-group AULC, while the highest-ranked (Harrier, 69.0) is the only one whose acquisition gain transfers to unseen events; Qwen’s large in-domain gain (+0.089 critical-recall AULC) reverses out of domain (−0.014). We further show that Matryoshka retention is policy-dependent: truncating Qwen to 512 dimensions costs ≤ 0.008 AULC when measured with random sampling, the static-probe analogue, but up to 0.054 under BADGE—an underestimate of $12\times$. For EmbeddingGemma at 512 dimensions the gap reaches $31\times$. We argue that the unit of validation for a multilingual encoder should be the representation–acquisition pipeline, not the frozen benchmark score, and we provide a frozen confirmatory protocol for such audits.

1 Introduction

Practitioners choose a multilingual sentence encoder by reading a leaderboard. MTEB and

its massive multilingual extension have made encoders comparable at scale (Muennighoff et al., 2023; Enevoldsen et al., 2025), and the release notes of every recent open encoder lead with an aggregate multilingual mean (Zhang et al., 2025; Vera et al., 2025). The same reflex governs efficiency: Matryoshka Representation Learning (Kusupati et al., 2022) lets a 1,024-dimensional embedding be truncated to 128, and the retained quality is reported as probe accuracy at each prefix length.

Both practices share an assumption that is invisible because it is almost always true in benchmarking: *the evaluation data is fixed*. A probe is fit on a given training set and scored on a given test set. The encoder is a passive feature extractor.

That assumption breaks in the setting where multilingual encoders matter most. When a team adapts an encoder to a new domain, language, or crisis with a few hundred labels, they must first decide *which* few hundred examples to annotate. Standard practice is pool-based active learning: score the unlabeled pool, request labels for a batch, retrain, repeat. Every quantity in that score—predictive uncertainty, nearest-neighbor distance, gradient embedding—is computed *in the encoder’s representation space*. The encoder therefore selects its own training data, and a small change in its geometry propagates into a different labeled set, a different classifier, and a different final score (Figure 1).

This paper asks what happens to standard multilingual-encoder evaluation practice when that feedback edge is restored. We audit three massively multilingual open encoders at 300M–600M parameters—Harrier-OSS-v1-0.6B (~94 languages), Qwen3-Embedding-0.6B (100+), and EmbeddingGemma-300M (100+)—inside a controlled low-budget selection loop. We cross them with three representation treatments (frozen, a standard supervised-contrastive residual adapter,

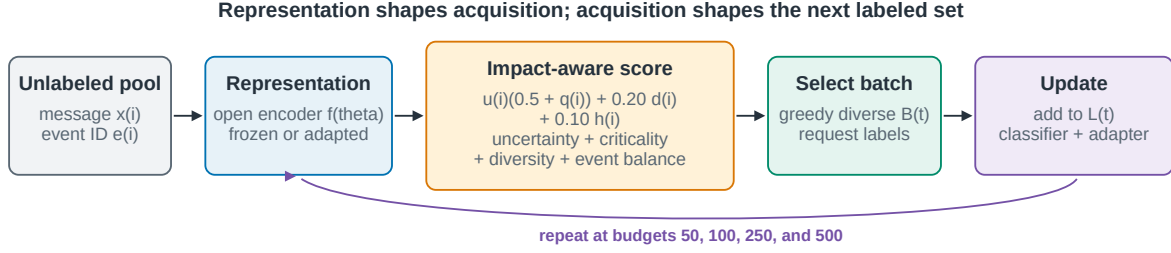


Figure 1: **The data-selection loop.** A multilingual encoder supplies both the uncertainty u_i and the geometry d_i that an acquisition policy uses to choose the next annotation batch; that batch becomes the supervision that trains the next classifier and optional adapter. Because the representation sits *inside* the loop, two encoders with near-identical benchmark scores can induce different labeled sets and therefore different downstream models. Static probe evaluation cuts this feedback edge and cannot observe the difference.

and a priority-weighted variant), five acquisition policies, five paired seeds, and budgets of 50–500 labels, for 975 configurations. Our testbed is humanitarian crisis text classification, a domain with genuinely scarce annotation capacity, explicit group structure (disaster events), and rare high-stakes categories—so that we can measure not only average quality but worst-group robustness and rare-class recall, the outcomes an aggregate benchmark mean hides.

We report three findings.

Benchmark rank does not predict adaptation rank. On MTEB multilingual the encoders order Harrier (69.0) > Qwen (64.33) > EmbeddingGemma (61.15). Inside the loop, this ordering does not hold and is not even stable across evaluation settings. EmbeddingGemma, the smallest and lowest-ranked, achieves the best held-out macro-F1 (0.425 AULC) and worst-group macro-F1 (0.423). Qwen attains the largest in-domain acquisition gain (+0.089 critical-recall AULC) but the gain *reverses* on unseen events (−0.014), whereas Harrier’s smaller in-domain gain (+0.058) is the one that transfers (+0.057).

Matryoshka retention is policy-dependent. Measured with random sampling—the closest analogue to a static probe, since the labeled set is then independent of the representation—512-dimensional Qwen is essentially lossless (≤ 0.008 AULC across three outcomes and two test settings), which is the number a conventional compression study would publish. Under an informed acquisition policy the same truncation costs up to 0.054 AULC in critical recall, a $12\times$ underestimate; for EmbeddingGemma at 512 dimensions the ratio reaches $31\times$. Truncation perturbs the geometry and the probabilities that selection depends

on, and a random-sampling probe is constitutionally unable to detect this.

Low-rank adaptation of multilingual encoders is unreliable at these budgets. Frozen representations win the majority of held-out outcomes, and a priority-weighted adapter objective is not consistently better than a standard one. We report this negative result in full.

Our contributions are: (i) a protocol for auditing multilingual encoders inside the data-selection loop, with the confirmatory analysis frozen before inspecting the held-out corpus; (ii) evidence that multilingual benchmark rank and Matryoshka probe retention both fail to predict behavior under low-budget adaptation, with the magnitude of the failure quantified; and (iii) a 975-configuration study, with negative results retained, that fits on a single 12 GB GPU slice.

We are explicit about scope: our two testbeds are English corpora. We evaluate *multilingual encoders*, not cross-lingual transfer. Section 5.3 states precisely what this does and does not license, and the Limitations section describes the multilingual extension this design was built to accept.

2 Related Work

Multilingual encoder evaluation. MTEB standardized embedding evaluation across tasks (Muennighoff et al., 2023), and MMTEB extended it to more than 500 tasks and 250 languages (Enevoldsen et al., 2025). These benchmarks are the de facto model-selection instrument for multilingual encoders. Their protocol is static: a fixed train/test split, a fixed probe or retrieval setup. Recent analyses question how far aggregate ranks transfer to specific deployments, partic-

ularly for under-represented languages (Uemura et al., 2026). We raise a structurally different objection: even holding language and domain fixed, a static protocol cannot observe an encoder’s effect on *which data gets labeled*, which is the dominant factor at small budgets.

Compact and truncatable representations.

Matryoshka Representation Learning trains nested subspaces so that prefixes remain usable (Kusupati et al., 2022). EmbeddingGemma exposes 768/512/256/128 prefixes (Vera et al., 2025) and Qwen3-Embedding supports user-defined dimensions from 32 to 1,024 (Zhang et al., 2025). Retention is conventionally validated with downstream probe accuracy at each prefix. We show this validation is optimistically biased whenever the deployed pipeline selects its own training data, and quantify the bias.

Language and domain adaptation. Adapter and residual methods adapt pretrained encoders cheaply (Houlsby et al., 2019; Pfeiffer et al., 2020), and supervised contrastive objectives are common for representation shaping (Khosla et al., 2020). Most evidence concerns budgets far above the 50–500 labels considered here. Our negative result—that frozen representations often win at these budgets—is a caution for the low-resource adaptation regime that motivates much multilingual work.

Active learning and its fragility. Acquisition functions select examples by uncertainty, by coverage (Sener and Savarese, 2018), or by both through gradient embeddings (Ash et al., 2020); see Settles (2009) for a survey. Pretrained-encoder studies show gains under small and imbalanced budgets (Ein-Dor et al., 2020), but Ghose and Nguyen (2024) demonstrate that rankings are highly sensitive to the surrounding pipeline, including the representation. Cost-aware acquisition for African NLP (Dossou et al., 2026) makes the related point that a uniform label count is the wrong objective in low-resource settings. We take the representation, rather than the acquisition function, as the object of study: for us, acquisition is the instrument that exposes representation differences a static probe cannot.

Humanitarian text classification as a testbed.

Crisis social-media corpora support classification of humanitarian information types across events

Encoder	Par.	Dim	Lang.	MTEB [†]	MRL
Harrier-OSS	0.6B	1024	~94	69.0	—
Qwen3-Emb.	0.6B	1024	100+	64.33	32–1024
Emb.Gemma	0.3B	768	100+	61.15	128–768

Table 1: Multilingual encoders under study. [†]Mean task score on the multilingual MTEB track as reported by each model’s own documentation; the benchmark revision is not identical across releases, so we treat these as indicative standing rather than a controlled comparison. “MRL” gives documented truncatable output dimensions; Harrier documents none, which is itself a reproducibility gap across otherwise comparable encoders.

(Imran et al., 2016). CrisisMMD provides annotations for seven disasters (Alam et al., 2018), Crisis-Bench consolidates heterogeneous sources (Alam et al., 2020), and HumAID provides fixed benchmarks over 19 events (Alam et al., 2021). Class imbalance and event shift are the recognized difficulties (Zou et al., 2023), and domain-adaptive pre-training has been applied to them (Seeberger and Riedhammer, 2022). We use this domain because it supplies group structure and rare consequential classes, not because we propose a crisis-specific method. Notably, although crisis streams are natively multilingual, the standard annotated benchmarks are English-only—a gap we return to in the Limitations section.

3 Encoders Under Study

Table 1 summarizes the three encoders. They are deliberately close in deployment cost (300M–600M parameters, all runnable on a 12 GB slice) and broad in language coverage (94–100+ languages), but they differ in reported multilingual benchmark standing by nearly eight points and in Matryoshka support.

Each message is encoded once and L2-normalized. Prompts follow each model’s documented convention and ask for a representation of a crisis-related social-media message for humanitarian information classification. Embeddings are cached in float16; downstream operations use float32. Because Harrier documents no truncatable prefixes, the compression analysis in Section 6.4 covers Qwen and EmbeddingGemma only.

4 The Data-Selection Loop

4.1 Setting

Let $U = \{(x_i, e_i)\}_{i=1}^N$ be an unlabeled pool of messages with group (event) identifiers, and L_t the labeled set after round t . Encoder f_θ maps x_i to a normalized vector z_i ; a classifier g_t trained on L_t produces $p_t(y | x_i)$. Each round, a policy selects $B_t \subset U \setminus L_t$ under a fixed total budget. The essential structural point is that g_t , and hence every acquisition score, is a function of f_θ .

Let $\mathcal{C}$ denote a set of rare, consequential labels. On CrisisMMD $\mathcal{C}$ is injured_or_dead_people and missing_or_found_people; HumAID adds requests_or_urgent_needs. These are task-level proxies for priority. We retain macro-F1 and add critical-category macro recall, worst-group (worst-event) macro-F1, and worst-group critical recall, so that encoder differences hidden by an aggregate mean become visible.

4.2 Representation Treatments

Frozen. The base vector $z = f(x)$, used without adaptation.

Standard residual adapter. For base embedding $x \in \mathbb{R}^d$, a rank- r adapter ($r = 64$) computes

$$z = \text{norm}(x + 0.1 W_{\text{up}} \text{GELU}(W_{\text{down}} x)), \quad (1)$$

where the small residual coefficient anchors the result to the pretrained geometry. At each budget the adapter is trained on the current labeled set for 100 epochs with cross-entropy plus $0.25 \times$ a supervised contrastive loss (Khosla et al., 2020); a class-balanced logistic regression is then fit on the adapted embeddings.

Priority-weighted adapter. This variant weights critical training examples by three, adds a binary critical/non-critical objective with coefficient 0.5, and a cosine-drift penalty with coefficient 0.1. It tests whether task priority should shape the representation itself. It does not consistently help (Section 6.3).

4.3 Acquisition Policies

We compare five policies. **Random** samples uniformly and is our static-probe analogue: it is the only policy whose labeled set is independent of the representation. **Entropy** takes the highest normalized predictive entropy u_i . **Core-set** performs farthest-first traversal (Sener and Savarese, 2018)

in a fixed 64-dimensional Gaussian random projection of the task representation. **BADGE** performs farthest-first selection over normalized hallucinated gradient embeddings (Ash et al., 2020). Core-set and BADGE are embedding-level approximations with a deterministic 10,000-example candidate cap; we do not claim exact reproduction of their original settings.

Priority-aware acquisition, our fifth policy, is included because it depends on the representation through *both* probabilities and geometry and therefore probes the loop most aggressively. With $q_i = \sum_{y \in \mathcal{C}} p_t(y | x_i)$, expected critical uncertainty is $b_i = u_i(0.5 + q_i)$; multiplication retains uncertainty as the learning signal rather than favoring already-confident examples. During sequential batch construction, diversity is $d_i = 1 - \max_{j \in L_t \cup B_t} \cos(z_i, z_j)$ and a group-balance bonus is $h_i = (1 + n_{e_i})^{-1/2}$, with n_{e_i} the count already selected from group e_i . The score is

$$s_i = b_i + 0.20 d_i + 0.10 h_i, \quad (2)$$

selected greedily with d_i and h_i updated after each pick. Group IDs are metadata, not labels.

5 Evaluation Protocol

5.1 Testbeds and Split Roles

Table 2 gives the two settings. The original HumAID event-wise split was used for method development and is not confirmatory, because intermediate results informed revisions. We then built an event-held-out HumAID protocol: nine events for training, the 2016 Ecuador earthquake for validation, and Cyclone Idai 2019, the Greece wildfires 2018, and the Puebla Mexico earthquake 2017 for test. CrisisMMD’s official split was left untouched until the encoders, treatments, outcomes, and comparisons were frozen (Appendix A). Duplicate text rows induced by repeated images were removed from CrisisMMD training, leaving 11,584 unique messages; no conflicting text labels or cross-split ID overlaps were found.

CrisisMMD’s test set contains 70 injured/dead and six missing/found messages; the held-out HumAID test contains 1,182 injured/dead, 39 missing/found, and 229 urgent-request messages. The very small CrisisMMD missing/found class is why we retain the larger held-out-event analysis as a counterweight. Hyperparameters were fixed during development; validation labels do not select among reported test configurations.

Setting	Train	Valid	Test	Ev. [†]	Cls.
CrisisMMD official	11,584	2,237	2,237	7/7/7	8
HumAID held-out	34,383	1,563	7,463	9/1/3	10

Table 2: Evaluation data. [†]Train/validation/test event counts. CrisisMMD is the untouched cross-corpus confirmation; the HumAID protocol tests unseen events. Both are English.

5.2 Design, Metrics, and Uncertainty

Each encoder is crossed with three representation treatments, five acquisition policies, and five seeds, at total budgets of 50, 100, 250, and 500. Within a dataset and seed all policies start from the same 50 uniformly sampled examples, so the first point is identical and later differences are paired. This yields 225 runs per dataset plus the development sweep and 300 compression runs: 975 configurations and 3,900 budget-level evaluations.

Primary outcomes are macro-F1, critical-category macro recall, and worst-event macro-F1. We also report worst-event critical recall, expected calibration error (10 bins), a binary critical Brier score, the critical fraction of the selected set, the minimum selected count across events, and Jensen–Shannon divergence (JSD) between the selected event distribution and the pool or a uniform target. For a metric measured at budgets (50, 100, 250, 500), normalized AULC is the trapezoidal integral divided by 500–50; AULC summarizes the whole low-budget trajectory rather than a single endpoint. Comparisons are paired by seed, bootstrapped 10,000 times over the five seed-wise differences, and reported as percentile 95% intervals. With five seeds these describe uncertainty; we do not present them as asymptotic significance tests.

All work runs on one 12 GB H100 MIG slice with five CPU cores. Extraction uses batch size 16 and maximum length 192; per-encoder cost is in Appendix C. All jobs completed without failures, timeouts, or out-of-memory errors.

5.3 Scope of the Multilingual Claim

Our object of study is a class of *massively multilingual* encoders, and our findings concern how their representational differences behave inside an adaptation loop. Both evaluation corpora are English. We therefore claim: the multilingual benchmark score that governs encoder selection, and the Matryoshka probe retention that governs compression

decisions, both fail to predict low-budget adaptation behavior—demonstrated on English data. We do *not* claim any result about cross-lingual transfer, about performance in any specific non-English language, or that the observed ordering of these encoders would persist in another language. Because the failure we identify is a property of the evaluation *protocol* rather than of English, we expect it to be at least as severe under language shift, where representation geometry differs far more—but that is a hypothesis this paper does not test.

6 Results

6.1 Benchmark Rank Does Not Predict Adaptation Rank

Table 3 reports the pre-specified paired comparison between priority-aware and random acquisition under the standard adapter, on both the untouched CrisisMMD confirmation and the event-held-out HumAID test.

Three observations follow. First, on the untouched CrisisMMD confirmation all three encoders improve critical-recall AULC with intervals excluding zero, so the loop is doing something real for every encoder. Second, the *size* of that benefit does not track multilingual benchmark standing: the top-ranked encoder (Harrier, 69.0) gains least, and the bottom-ranked (EmbeddingGemma, 61.15) gains nearly as much as Qwen. Third, and most consequentially, the ordering *inverts* under event shift. Harrier’s smaller in-domain gain is the only one that transfers to unseen events; Qwen’s much larger in-domain gain becomes slightly negative. An evaluation that reported only the confirmation set would have recommended precisely the wrong encoder for deployment on a new event.

The pattern persists when we look at absolute winners rather than deltas. On held-out events, the best critical-recall AULC belongs to *frozen* Qwen with entropy (0.515), while the best macro-F1 (0.425) and best worst-event macro-F1 (0.423) both belong to *frozen EmbeddingGemma with random sampling*—the smallest, lowest-ranked encoder, unadapted, with no acquisition policy at all. On CrisisMMD the best critical recall is frozen EmbeddingGemma with entropy (0.429). No encoder is best at everything, and the benchmark leader is not best at anything we measure.

For standard Qwen on CrisisMMD, priority-aware acquisition also beats entropy by 0.0575 [0.0108, 0.1088] in critical recall and 0.0391

Setting	Encoder (MTEB)	Δ Critical recall	Δ Macro-F1	Δ Worst-event F1
CrisisMMD (untouched)	Qwen (64.33)	+0.0885 [.0585, .1177]	+0.0261 [.0078, .0406]	+0.0324 [.0011, .0552]
	EmbeddingGemma (61.15)	+0.0870 [.0401, .1296]	+0.0189 [.0048, .0332]	+0.0181 [−.0047, .0341]
	Harrier (69.0)	+0.0575 [.0210, .0897]	+0.0156 [−.0022, .0368]	+0.0207 [.0053, .0419]
HumAID (unseen events)	Harrier (69.0)	+0.0571 [.0123, .0933]	+0.0048 [−.0075, .0169]	+0.0075 [−.0061, .0247]
	EmbeddingGemma (61.15)	+0.0424 [−.0252, .1079]	−.0097 [−.0263, .0063]	−.0062 [−.0220, .0095]
	Qwen (64.33)	−.0141 [−.0518, .0313]	−.0025 [−.0257, .0283]	−.0027 [−.0315, .0272]

Table 3: Priority-aware minus random acquisition under the standard adapter, paired by seed; brackets are descriptive 95% bootstrap intervals and bold indicates an interval excluding zero. Parentheses give each encoder’s multilingual MTEB standing. The in-domain ordering (Qwen > EmbeddingGemma > Harrier) is neither the benchmark ordering nor the out-of-domain ordering: under event shift the ranking inverts, and Harrier is the only encoder whose gain retains an interval excluding zero.

[0.0312, 0.0470] in worst-event macro-F1. Against BADGE the differences are uncertain and BADGE has the stronger final-budget macro-F1. Figure 2 shows the trade-off: at 500 labels standard Qwen reaches 0.446 macro-F1 / 0.460 critical recall / 0.383 worst-event F1 with priority-aware acquisition, versus 0.468 / 0.391 / 0.374 with BADGE. These are distinct Pareto points, not a ranking.

6.2 What the Loop Changes in the Labeled Data

The encoder’s influence is legible in the labeled set itself. At 500 CrisisMMD labels, random sampling selects 3.36% critical examples. With the standard adapter, priority-aware acquisition raises this to 26.76% for Qwen, 34.16% for EmbeddingGemma, and 27.72% for Harrier—an eight-to ten-fold change in what a human annotator is shown, driven entirely by the representation and the policy. Qwen’s minimum selected count per event rises from 16.2 to 39.6 and its JSD from a uniform event allocation falls from 0.0921 to 0.0263 (0.0191 for EmbeddingGemma, 0.0256 for Harrier). Figure 4 in Appendix D visualizes this composition.

This is the mechanism behind Section 6.1: the encoder does not merely score a fixed test set, it determines the training distribution. It also explains why critical recall and macro-F1 need not move together, since the selected distribution deliberately departs from the naturally imbalanced pool.

Calibration effects are mixed. On CrisisMMD the binary critical Brier score improves at 500 labels for all three standard-adapted encoders (e.g., 0.0265 to 0.0145 for Qwen), but expected calibration error changes little. On held-out HumAID, critical Brier improves slightly for Harrier and EmbeddingGemma and worsens slightly for Qwen.

We make no general calibration claim.

6.3 Low-Rank Adaptation Is Not Uniformly Helpful

Priority weighting inside the residual adapter is not consistently better than the standard adapter and is sometimes worse, across policies and both testbeds. Even standard adaptation is not universally beneficial: as noted above, frozen configurations lead all three held-out outcomes. At 50–500 labels there is simply not enough signal to reshape a well-trained multilingual representation reliably, and the pretrained geometry is a strong prior worth keeping. Our positive result is therefore about co-designing acquisition with a *conservative* representation, not about a superior adaptation objective. For a community that routinely proposes adapters for low-resource languages, the frozen baseline deserves to be run.

6.4 Matryoshka Retention Is Policy-Dependent

We truncate Qwen and EmbeddingGemma to 128, 256, and 512 prefix dimensions and renormalize. Under random sampling—where the labeled set does not depend on the representation, making this the closest analogue to a static probe evaluation—retention is excellent and matches what compression studies report. Across three outcomes and both test settings, 512-dimensional Qwen stays within 0.008 AULC of full dimension at half the storage; 512-dimensional EmbeddingGemma stays within 0.014 at two-thirds; and even 128-dimensional Qwen, an 87.5% storage reduction, stays within 0.030.

Table 4 shows what that number misses. For each truncation we contrast the random-sampling estimate of critical-recall degradation with the

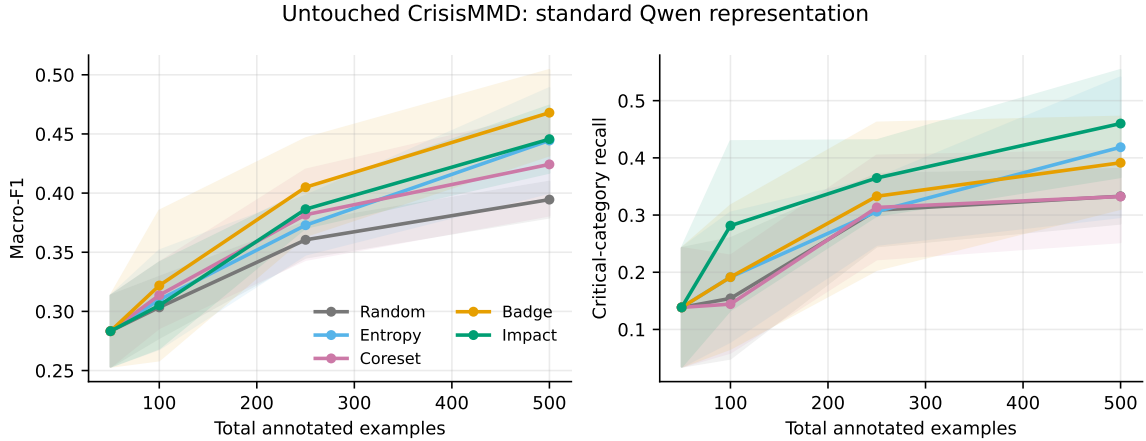


Figure 2: CrisisMMD learning curves for the standard-adapted Qwen encoder; shading is one standard deviation over five paired seeds. All policies share the same 50-label initialization within a seed. The curves separate only because the encoder scores the pool differently under each policy.

worst degradation observed across the four informed policies.

The gap is large and systematic. At 512 dimensions—the setting a practitioner would most readily adopt, since the probe reports it as free—Qwen loses 0.054 critical-recall AULC under BADGE against a random-sampling estimate of 0.004, a $12\times$ underestimate, and EmbeddingGemma loses 0.076 under entropy where the probe reports a slight *improvement*. The pathological ratios ($251\times$) arise where the random estimate is near zero, and should be read as “the probe sees nothing” rather than as a precise multiplier; the absolute degradations are the meaningful quantity, and they reach 0.110 AULC. The effect is not confined to one corpus: on unseen events, 128-dimensional EmbeddingGemma loses 0.122 critical-recall AULC under priority-aware acquisition versus 0.023 under random.

The mechanism is direct. Truncation preserves enough of the subspace for a fixed probe to separate classes, which is exactly what Matryoshka training optimizes. It does not preserve the finer nearest-neighbor structure that Coreset and BADGE traverse, nor the calibrated tail probabilities that entropy and priority-aware scoring depend on. Degraded selection compounds: a worse batch at budget 100 yields a worse classifier that selects a worse batch at 250. Figure 3 shows the frozen random-sampling retention curves that a conventional study would publish—smooth and reassuring—which Table 4 should be read against.

The practical recommendation is conditional rather than a single number: 512 dimensions are

safe when the downstream pipeline does not select its own data, and full dimensionality should be retained when it does. More generally, compression claims for multilingual encoders should state the acquisition regime under which retention was measured.

7 Discussion

The unit of validation is the pipeline, not the encoder. Every result above stems from one structural fact: at small budgets, the encoder chooses its own training data. A benchmark mean and a prefix-retention curve both hold the data fixed and therefore measure a system that is not the one deployed. The gap is not small—it inverted an encoder ranking under distribution shift and hid up to an order of magnitude of compression damage. We think multilingual encoder cards should report at least one representation-dependent selection outcome alongside static scores.

Aggregate scores hide the outcomes that matter under shift. A single multilingual mean cannot express that the highest-ranked encoder is the most robust to event shift while the lowest-ranked is the best frozen performer. This is the familiar objection to aggregate leaderboards, but the selection loop sharpens it: the disagreement here is not between two static metrics, it is between a static metric and the encoder’s behavior as an active component. Worst-group and rare-class outcomes made the disagreement visible; macro averages alone would not have.

Encoder	Dim	Store	CrisisMMD			HumAID unseen events	
			Random	Worst policy	Ratio	Random	Worst policy
Qwen	512	50.0%	−0.004	−0.054 (BADGE)	12.4×	−0.001	−0.037 (Priority)
	256	25.0%	−0.013	−0.075 (Priority)	5.7×	+0.026	−0.030 (Priority)
	128	12.5%	−0.007	−0.053 (Priority)	7.2×	−0.010	−0.039 (Priority)
Embedding-Gemma	512	66.7%	+0.002	−0.076 (Entropy)	31.7×	+0.002	−0.023 (Priority)
	256	33.3%	+0.007	−0.110 (Entropy)	16.4×	−0.008	−0.059 (Priority)
	128	16.7%	+0.000	−0.080 (Entropy)	251×	−0.023	−0.122 (Priority)

Table 4: Matryoshka retention for critical-recall AULC, as change from the full-dimensional frozen representation. “Random” is the static-probe analogue, in which the labeled set is independent of the representation; “Worst policy” is the largest degradation over entropy, Core-set, BADGE, and priority-aware acquisition, with the responsible policy named. “Ratio” is the CrisisMMD underestimate factor. A conventional compression study reports the “Random” column and would conclude that all six truncations are safe.

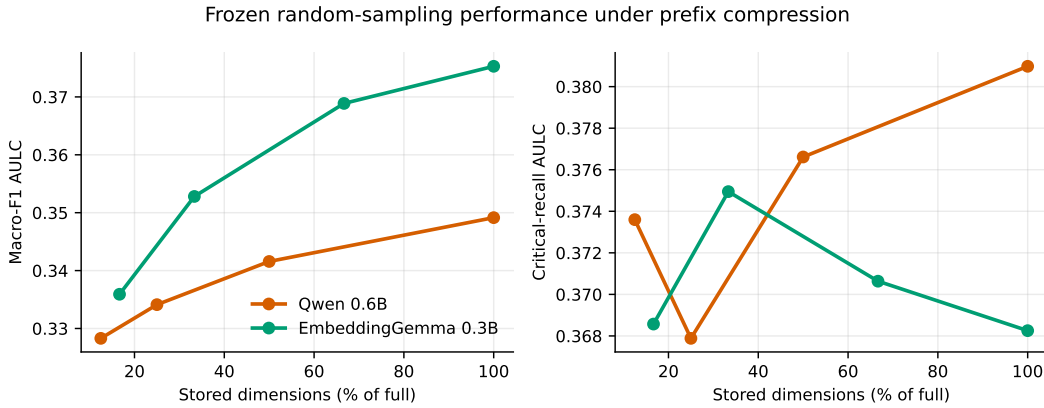


Figure 3: CrisisMMD AULC under prefix compression with frozen representations and random sampling; the x-axis is stored dimensions as a percentage of full output. This is the view a static compression study produces, and it is reassuring. Table 4 shows that the same truncations degrade representation-dependent acquisition by up to an order of magnitude more.

Frozen is a serious baseline in low-resource adaptation. Frozen representations led every held-out outcome we measured. Given how much multilingual work targets exactly this budget regime for under-resourced languages, we read this as a call to report the frozen baseline rather than assume adaptation helps.

Domain caution. Our testbed involves crisis messages about injury, death, and urgent need. The priority taxonomy is a transparent task proxy, not a validated operational instrument. A real deployment should define priorities with affected communities and practitioners, audit false negatives, and retain human control over consequential decisions. Group balancing across events is not a substitute for geographic, linguistic, or demographic equity.

8 Conclusion

We audited three massively multilingual sentence encoders inside the data-selection loop that governs low-budget adaptation, across 975 configurations with a confirmatory protocol frozen in advance. Multilingual benchmark rank did not predict adaptation behavior: the ordering of encoders changed between in-domain and unseen-event evaluation, and the lowest-ranked encoder was the strongest frozen performer. Matryoshka prefix retention, validated the conventional way, understated degradation under representation-dependent acquisition by up to an order of magnitude. Low-rank adaptation at 50–500 labels was unreliable, and frozen representations were a strong baseline throughout. These are evaluation-protocol findings rather than claims about any particular encoder, and the protocol they call for—measure the representation–acquisition pipeline, not the frozen score—extends naturally to the cross-lingual set-

ting we could not test here.

Limitations

English evaluation of multilingual encoders.

This is the central limitation. Our encoders span 94–100+ languages but both testbeds are English, because the standard annotated humanitarian benchmarks are English-only even though crisis streams are natively multilingual. We consequently cannot claim anything about cross-lingual transfer, about behavior in any specific non-English language, or that the encoder ordering we observe would persist under language shift. The design was built to accept that extension: the loop treats the group identifier e_i as opaque metadata, so replacing disaster events with languages requires no methodological change. The natural next study runs this protocol on a multilingual humanitarian corpus such as HumSet (Fekih et al., 2022), whose English, French, and Spanish documents would let group balance be measured over languages and would test whether benchmark rank predicts *cross-lingual* adaptation any better than it predicts within-language adaptation. We expect the failure to be more severe under language shift, but we did not test it.

Other limitations. CrisisMMD’s test split contains only six missing/found examples, so its two-class critical macro recall is sensitive to a handful of cases; the larger HumAID critical categories are an essential counterweight. Our priority labels and group-balance objective are proxies for operational utility; we did not measure response time, resource allocation, or outcomes for affected people, and no humanitarian practitioners evaluated the selected messages. The loop assumes group metadata is available, and a deployment without it would need inferred strata. Annotation costs are treated as uniform, and we do not model annotator disagreement, expertise, or trauma exposure. Our Core-set and BADGE baselines are embedding-level approximations with a candidate cap. Five paired seeds support transparent within-seed comparison but not asymptotic significance claims. The compression analysis covers two encoders because Harrier documents no truncatable prefixes. The acquisition coefficients are fixed heuristics; component and sensitivity ablations remain valuable follow-up work. Finally, the MTEB figures in Table 1 come from each model’s own documentation and are not a controlled re-evaluation, so we

use them as indicative standing rather than as a precise ranking.

Ethics Statement

The study uses previously released crisis datasets and does not redistribute message text. Crisis posts may contain sensitive information about injuries, deaths, locations, or requests for aid. Models trained on such data can amplify misinformation, expose individuals, or cause responders to overlook communities whose language and connectivity are underrepresented. The system prioritizes examples for human annotation; it does not autonomously determine aid eligibility or operational response.

A priority-aware policy creates both benefits and risks. It can direct scarce annotation toward rare urgent categories, but an incorrect criticality model can repeatedly surface sensational or ambiguous content. Group balancing can reduce dominance by large events, yet equal allocation may be inappropriate when severity and population exposure differ. Any deployment should log selection scores, monitor per-group false negatives, support priority overrides, and establish escalation paths for uncertain critical cases. Human annotators should receive appropriate support and should not be exposed to distressing content without informed workplace safeguards.

Our claim is deliberately narrow: the evaluation protocol we criticize mismeasures encoder behavior, and the policy we study changes which examples receive labels. Neither establishes a positive societal outcome, which would require participatory priority definition, practitioner evaluation, and prospective deployment study.

References

- Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. [CrisisMMD: Multimodal Twitter datasets from natural disasters](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.
- Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. [HumAID: Human-annotated disaster incidents data from Twitter with deep learning benchmarks](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 933–942.
- Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2020. [CrisisBench: Benchmarking](#)

- crisis-related social media datasets for humanitarian information processing. *arXiv preprint arXiv:2004.06774*.
- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. [Deep batch active learning by diverse, uncertain gradient lower bounds](#). In *International Conference on Learning Representations*.
- Bonaventure F. P. Dossou, Ines Arous, Audrey Durand, and Jackie Chi Kit Cheung. 2026. [Active learning with non-uniform costs for African natural language processing](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6644–6656.
- Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. [Active learning for BERT: An empirical study](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7949–7962.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, et al. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Selim Fekih, Nicolò Tamagnone, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, and Navid Rekabsaz. 2022. [HumSet: Dataset of multilingual information extraction and classification for humanitarian crisis response](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4379–4389.
- Abhishek Ghose and Emma Thuong Nguyen. 2024. [On the fragility of active learners for text classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22217–22233.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.
- Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. [Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1638–1643.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35.
- Niklas Muennighoff, Nouamane Tazi, Loïc 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.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. [MAD-X: An adapter-based framework for multi-task cross-lingual transfer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7654–7673.
- Philipp Seeberger and Korbinian Riedhammer. 2022. [Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning](#). In *Proceedings of the Second Workshop on NLP for Positive Impact*, pages 70–78.
- Ozan Sener and Silvio Savarese. 2018. [Active learning for convolutional neural networks: A core-set approach](#). In *International Conference on Learning Representations*.
- Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison.
- 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 2026 Conference of the European Chapter of the Association for Computational Linguistics*.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panam, et al. 2025. [EmbeddingGemma: Powerful and lightweight text representations](#). *arXiv preprint arXiv:2509.20354*.
- 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](#). *arXiv preprint arXiv:2506.05176*.
- Henry Peng Zou, Yue Zhou, Weizhi Zhang, and Cornelia Caragea. 2023. [DeCrisisMB: Debaised semi-supervised learning for crisis tweet classification via memory bank](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6104–6115.

A Frozen Confirmatory Protocol

Before inspecting CrisisMMD model results we fixed: the three encoders; the frozen, standard, and priority-weighted treatments; five acquisition policies; five seeds; budgets 50, 100, 250, and 500; primary outcomes; the candidate cap; and paired-bootstrap comparisons. We also fixed the interpretation rule that no configuration would be called dominant unless critical recall and macro-F1 both improved without a material worst-event loss. The original HumAID split was declared development-only, event-held-out HumAID a robustness test, and CrisisMMD the untouched confirmation.

B Hyperparameters

Parameter	Value
Adapter rank	64
Adapter residual scale	0.1
Adapter epochs per budget	100
Adapter learning rate	2×10^{-3}
Adapter weight decay	10^{-4}
Supervised contrastive coefficient	0.25
Critical sample weight	3.0
Critical BCE coefficient	0.5
Representation drift coefficient	0.1
Selection projection dimension	64
Candidate cap	10,000
Diversity coefficient	0.20
Group-balance coefficient	0.10
Logistic regression C	1.0
Bootstrap resamples	10,000

Table 5: Fixed experimental hyperparameters.

C Encoding Cost

Encoder	Dim.	Seconds	Peak GiB
Harrier 0.6B	1024	76.3	1.18
Qwen 0.6B	1024	77.2	1.36
EmbeddingGemma 0.3B	768	73.4	0.67

Table 6: Measured extraction on the 11,584-item CrisisMMD training split, one H100 MIG slice, batch size 16, maximum length 192. All encoders use well under the 12 GB allocation.

D Selected-Set Composition

E Additional Reproducibility Details

Each result JSON records the dataset, encoder, full or truncated dimension, representation treatment, policy, seed, budget sequence, labels, critical labels, and all scalar and per-event metrics. Aggregation searches result directories recursively, computes normalized trapezoidal AULC per run, and pairs methods by seed before bootstrapping.

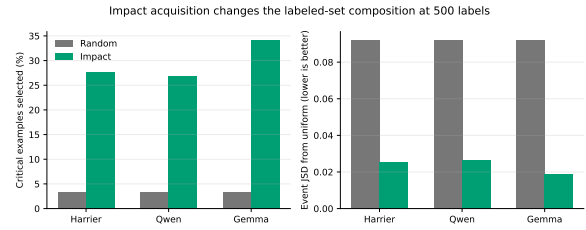


Figure 4: Composition of the standard-representation CrisisMMD labeled set at 500 labels. Priority-aware acquisition selects substantially more critical examples and allocates annotations closer to uniform across events (lower JSD).

Dataset preparation records source URLs, archive checksums, row counts, duplicate removals, class counts, and event counts. Embedding manifests record model identifiers, prompts, batch size, maximum length, stored dtype, dimensions, split run-times, and peak allocated GPU memory.