

Who Gets Annotated? Impact-Aware Active Learning for Robust Humanitarian Text Classification

Divyajot Singh

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

Abstract

Humanitarian organizations must adapt text classifiers to new crises with small annotation budgets, yet conventional active learning optimizes model uncertainty without representing operational criticality or coverage across disaster events. We study an impact-aware batch acquisition policy that combines expected criticality, uncertainty, representation-space diversity, and event balance. Our evaluation crosses three recent open embedding backbones, three representation treatments, five acquisition strategies, five paired seeds, and total budgets of 50–500 labels. We separate exploratory development, event-held-out robustness on HumAID, and an untouched cross-dataset confirmation on CrisisMMD. On CrisisMMD, impact-aware selection with a conservative residual adapter improves Qwen’s area under the learning curve (AULC) over random sampling by 0.089 for critical-category recall, 0.026 for macro-F1, and 0.032 for worst-event macro-F1. It raises the critical share of the 500-example labeled set from 3.4% to 26.8% and allocates annotations more evenly across events. Critical-recall gains also hold for Harrier and EmbeddingGemma on CrisisMMD and transfer to unseen HumAID events for Harrier, but not Qwen, exposing a consequential representation–dataset interaction. No configuration dominates every operational objective. Finally, under frozen random sampling, 512-dimensional Qwen halves embedding storage while remaining within 0.009 AULC of its full representation for macro-F1, critical recall, and worst-event F1 across both test settings. These results support impact-aware acquisition as a useful, but not always-on, policy and motivate reporting Pareto trade-offs rather than a single aggregate score.

1 Introduction

Social media can provide timely reports of casualties, urgent needs, infrastructure damage, and response efforts during disasters (Imran et al., 2016). Turning this stream into situational awareness requires classifiers that can distinguish multiple humanitarian information types. The events for which such systems are most needed, however, are also the events for which labeled data are initially scarce. Annotation capacity is limited, incoming distributions differ across disasters, and operationally critical categories are often rare.

Pool-based active learning (AL) appears well matched to this setting: a model selects a small batch of messages for annotation, learns from those labels, and repeats. Standard acquisition functions select uncertain or geometrically diverse examples. These criteria optimize generic learning value, not necessarily humanitarian value. An uncertainty sampler can spend a large fraction of its budget refining common classes, while a diversity sampler can under-cover a small disaster event. In a capacity-constrained response, the composition of the labeled set matters in addition to its average predictive performance.

Recent work also cautions that AL gains are fragile. Rankings change with the dataset, representation, classifier, and labeling budget, and sophisticated methods frequently fail to beat random sampling (Ghose and Nguyen, 2024). This is particularly important in crisis informatics, where event shifts and class imbalance already produce uneven class performance (Zou et al., 2023). A credible evaluation therefore needs more than one representation, one dataset, or one final-budget score.

We ask: *Can a lightweight acquisition policy preferentially discover operationally critical cases and improve coverage across disaster events without unacceptable losses in overall performance?*

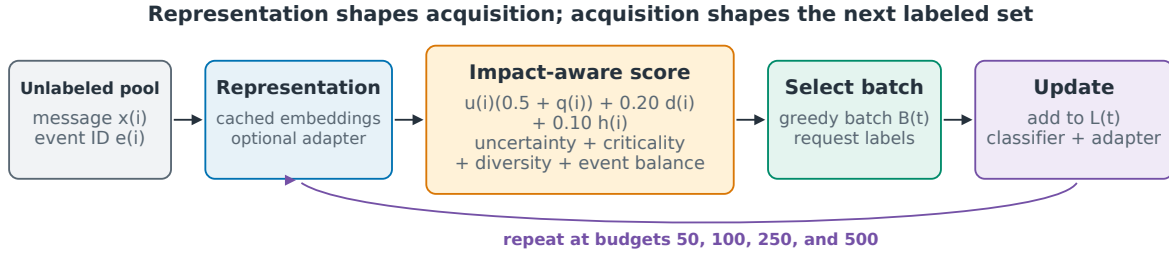


Figure 1: **IMPACTAL framework.** The encoder embeds messages once. Each round uses cached or adapted vectors to combine uncertainty (u_i), predicted criticality (q_i), representation diversity (d_i), and an event-balance bonus (h_i), requests labels for a greedily selected batch, and updates the classifier and optional adapter. The feedback loop motivates evaluating representation and acquisition jointly.

We introduce IMPACTAL, an impact-aware acquisition score over pretrained text embeddings. It favors examples that are both uncertain and likely to belong to designated critical classes, then adds representation-space diversity and a diminishing bonus for underrepresented events. We study it with frozen representations and two residual adapter variants. Rather than assume that representation adaptation is uniformly beneficial, we treat the interaction between representation and acquisition as an empirical question. Figure 1 makes this coupling explicit: the representation determines the uncertainty and geometry used for selection, and each selected batch changes the supervision used in the next round.

Our study has five distinguishing features. First, it includes recent open embedding families at 300M–600M parameters rather than a single encoder. Second, it compares frozen, standard task-adapted, and impact-weighted representations. Third, it evaluates random, entropy, Core-set, BADGE, and impact-aware acquisition at four small budgets. Fourth, it reports critical recall, macro-F1, worst-event performance, calibration, and selected-set composition over five paired seeds. Fifth, it separates a development split from an event-held-out HumAID test and an untouched CrisisMMD confirmation; the confirmatory protocol was frozen before inspecting CrisisMMD model results.

Our contributions are:

- an impact-aware batch acquisition policy that combines expected critical uncertainty, representation diversity, and event balance;
- a 975-configuration study of acquisition–representation interactions under humanitarian annotation budgets, including an

untouched cross-dataset confirmation and unseen-event robustness test;

- evidence that impact acquisition improves critical-recall AULC for all three standard-adapted backbones on CrisisMMD, while its cross-event transfer depends on the representation;
- a mechanism analysis showing how the policy changes the labeled set, together with negative evidence for impact-weighted representation adaptation; and
- measured compute and embedding-compression results demonstrating that the complete workflow fits on a 12 GB H100 MIG slice.

2 Related Work

Active learning for text classification. AL selects examples whose labels are expected to be most useful (Settles, 2009). Entropy sampling prioritizes predictive uncertainty; Core-set methods use farthest-first selection to cover the representation space (Sener and Savarese, 2018); and BADGE combines uncertainty and diversity through hallucinated gradient embeddings (Ash et al., 2020). Large empirical studies with pretrained language models show that AL can help under small and imbalanced labeling regimes (Eindor et al., 2020), but more recent evidence demonstrates that gains are highly sensitive to the surrounding prediction pipeline (Ghose and Nguyen, 2024). Cost-aware AL further shows that the acquisition objective should reflect constraints beyond a uniform label count (Dossou et al., 2026). We share that operational motivation, but model priority categories and event allocation rather than annotation time.

Humanitarian information processing. Crisis-related social-media corpora have enabled classification of humanitarian information types across many disaster events (Imran et al., 2016). CrisisMMD provides aligned text and image annotations for seven disasters (Alam et al., 2018); we use its text task. CrisisBench consolidated heterogeneous crisis datasets to support comparable benchmarking (Alam et al., 2020), while HumAID introduced a larger, deduplicated collection with fixed benchmarks across 19 events (Alam et al., 2021). These resources primarily evaluate models after labels are available. Our focus is which messages should receive the next scarce labels.

Class imbalance and event shift are central concerns. DeCrisisMB studies class-biased semi-supervised learning and out-of-distribution crisis classification (Zou et al., 2023). Entity-masked language modeling and multi-task learning can reduce event overfitting (Seeberger and Riedhammer, 2022). We complement these approaches by intervening on the labeled-set composition and measuring the worst event directly.

Efficient text representations. We evaluate Qwen3-Embedding-0.6B (Zhang et al., 2025), EmbeddingGemma-300M (Vera et al., 2025), and Harrier-OSS-v1-0.6B (Microsoft, 2026). These models provide dense representations; Qwen and EmbeddingGemma support truncated prefixes. Matryoshka-style representations motivate evaluating whether shorter prefixes retain task quality (Kusupati et al., 2022). Our compression analysis is coupled to acquisition: a representation that preserves a random linear probe need not preserve the geometry used by an impact-aware selector.

3 Problem Setting

Let $U = \{(x_i, e_i)\}_{i=1}^N$ be an unlabeled pool of messages and event identifiers, and let L_t be the labeled set after round t . A representation f_θ maps message x_i to a normalized vector z_i . A classifier g_t trained on L_t produces class probabilities $p_t(y | x_i)$. At each round, an acquisition policy selects a batch $B_t \subset U \setminus L_t$ under a total annotation budget.

Let $\mathcal{C}$ denote the set of operationally critical labels. On CrisisMMD, $\mathcal{C}$ contains `injured_or_dead_people` and `missing_or_found_people`. HumAID additionally includes `requests_or_urgent_needs`. These categories are task-level proxies for priority,

not a claim that every message within them has equal downstream utility. We therefore retain conventional macro-F1 and add critical-category macro recall, worst-event macro-F1, and worst-event critical recall.

The desired policy should discover critical examples, teach a useful classifier, and avoid concentrating the label budget in a few events. These goals can conflict. We consequently do not reduce them to a post-hoc weighted evaluation score; results are interpreted as Pareto trade-offs.

3.1 Impact Measurement Scope

We distinguish three levels in the proposed measurement chain. The *intervention* is the allocation of a fixed annotation budget. Its *proximal outcomes* are observable properties of the labeled set: the fraction of critical examples, minimum event coverage, and event-distribution divergence. Its *model outcomes* are critical recall, macro-F1, and worst-event performance on held-out messages. We measure all three levels. We do *not* measure downstream response decisions or societal outcomes. Accordingly, “impact-aware” describes which task-level priorities enter the acquisition policy; it does not assert that a benchmark gain itself constitutes positive humanitarian impact. Section 7 discusses the additional participatory and prospective evidence needed to establish that link.

4 Method

4.1 Representations

We encode each message once with an open pretrained embedding model and L2-normalize the result. Qwen and Harrier produce 1,024-dimensional embeddings; EmbeddingGemma produces 768 dimensions. Qwen and Harrier use a crisis-classification instruction; EmbeddingGemma uses its classification prefix. Embeddings are cached in float16, while downstream operations use float32.

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

Standard residual adapter. Given base embedding $x \in \mathbb{R}^d$, a rank- r adapter computes

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

where $r = 64$. The small residual coefficient anchors the representation to the pretrained geometry. At each budget we initialize a new adapter

and train it on the currently 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 classifier is then fit on the adapted labeled embeddings.

Impact-weighted adapter. This variant weights critical training examples by three and adds a binary critical/not-critical objective with coefficient 0.5 plus a cosine-drift penalty with coefficient 0.1. We include it to test whether operational priority should shape the representation as well as acquisition. It does not consistently improve performance (Section 6.4), an important negative result.

4.2 Impact-Aware Acquisition

For each unlabeled example, normalized predictive entropy is

$$u_i = -\frac{\sum_y p_t(y | x_i) \log p_t(y | x_i)}{\log |\mathcal{Y}|}, \quad (2)$$

and predicted criticality is $q_i = \sum_{y \in \mathcal{C}} p_t(y | x_i)$. We first define expected critical uncertainty

$$b_i = u_i(0.5 + q_i). \quad (3)$$

Multiplication retains uncertainty as the learning signal while increasing the value of plausibly critical examples. Adding criticality alone would favor already-confident examples that contribute little new information.

We compute acquisition geometry with normalized projected vectors $v_i = \text{norm}(Rz_i)$, where R is a fixed 64-dimensional Gaussian random projection with seed 17. During sequential batch construction, diversity is

$$d_i = 1 - \max_{j \in L_t \cup B_t} v_i^\top v_j, \quad (4)$$

and the event bonus is

$$h_i = \frac{1}{\sqrt{1 + n_{e_i}}}, \quad (5)$$

where n_{e_i} counts examples from event e_i in $L_t \cup B_t$. The final score is

$$s_i = b_i + 0.20d_i + 0.10h_i. \quad (6)$$

We take the top 10,000 candidates under b_i when the pool is larger, then greedily select the highest-scoring item and update d_i and h_i . Event IDs are metadata, not labels, and are available in both benchmark collections.

5 Experimental Design

5.1 Datasets and Split Roles

Table 1 summarizes the two evaluation settings. The original HumAID event-wise split was used for method development and is not treated as confirmatory because intermediate test results informed revisions. We then created an event-held-out HumAID protocol: nine events for training, the 2016 Ecuador earthquake for validation, and Cyclone Idai 2019, Greece wildfires 2018, and the Puebla Mexico earthquake 2017 for test. CrisisMMD’s official split was left untouched until the method, outcomes, and comparisons were frozen. Duplicate text rows induced by repeated images were removed from the CrisisMMD training split, 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 extremely small CrisisMMD missing/found class motivates retaining the larger held-out-event robustness analysis. Hyperparameters were fixed from development; validation labels are not used to select among the reported test configurations.

5.2 Models, Baselines, and Budgets

We use Harrier-OSS-v1-0.6B, Qwen3-Embedding-0.6B, and EmbeddingGemma-300M. Each is crossed with frozen, standard-adapted, and impact-weighted representations and five acquisition strategies:

1. **Random:** uniform sampling without replacement;
2. **Entropy:** highest normalized predictive entropy;
3. **Core-set:** farthest-first traversal in the 64-dimensional selection view, using an uncertainty-filtered candidate set;
4. **BADGE:** farthest-first selection over normalized hallucinated gradient embeddings (Ash et al., 2020); and
5. **Impact:** Equation 6.

The Core-set and BADGE implementations are embedding-level approximations with a deterministic 10,000-example candidate cap; we do not

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

Table 1: Evaluation data. [†]Train/validation/test event counts. CrisisMMD is the untouched cross-dataset confirmation; HumAID tests unseen events.

claim exact end-to-end reproduction of their original vision settings.

Total label budgets are 50, 100, 250, and 500. For a given dataset and seed, all strategies begin from the same 50 uniformly sampled examples. Within each backbone and representation, the first point is identical across strategies and subsequent differences are paired. We run five seeds. This yields 225 runs per dataset (3 models $\times$ 3 representations $\times$ 5 strategies $\times$ 5 seeds), plus the development sweep.

5.3 Metrics and Uncertainty

The frozen protocol specifies critical-recall AULC, macro-F1 AULC, critical recall at 500 labels, and worst-event macro-F1 at 500 labels as primary outcomes. We additionally summarize worst-event macro-F1 by AULC. Per-event macro-F1 averages only classes present in that event’s test set; worst-event F1 is the minimum over events. We also report worst-event critical recall, expected calibration error (10 bins), a binary critical/non-critical Brier score, the critical fraction of selected examples, the minimum selected count across events, and Jensen–Shannon divergence (JSD) between the selected event distribution and either the pool or a uniform target.

For metric m measured at budgets $B = (50, 100, 250, 500)$, normalized AULC is the trapezoidal integral divided by $500 - 50$. Comparisons are paired by seed. We bootstrap the five seed-wise differences 10,000 times and report percentile 95% intervals. With only five seeds, these intervals describe uncertainty; we do not present them as definitive asymptotic significance tests.

5.4 Compute and Reproducibility

All work runs on one 12 GB H100 MIG slice with five CPU cores and 31 GB host memory per job. Embedding extraction uses batch size 16 and maximum length 192. The full study contains 675 main configurations and 300 compression configurations, totaling 3,900 budget-level evaluations. All jobs completed without failures, timeouts, or

out-of-memory errors. Code records dataset audits, embedding manifests, per-run JSON, aggregate CSV tables, and fixed random seeds.

6 Results

6.1 Untouched Cross-Dataset Confirmation

Table 2 reports the pre-specified standard-adaptation comparison between impact and random acquisition on CrisisMMD. Critical-recall AULC improves for all three backbones, with entirely positive paired-bootstrap intervals. Qwen also improves macro-F1 and worst-event macro-F1 AULC. Its gain over random occurs in four of five seeds for macro-F1 and worst-event F1 and in all five seeds for critical recall.

For standard Qwen, impact also improves critical-recall AULC over entropy by 0.0575 [0.0108, 0.1088] and worst-event macro-F1 AULC by 0.0391 [0.0312, 0.0470]. Its macro-F1 difference from entropy is small and uncertain. Against BADGE, the critical-recall and worst-event differences are uncertain and BADGE has the stronger final-budget macro-F1. The result is therefore not universal dominance over strong batch AL baselines.

Figure 2 makes the trade-off visible. At 500 labels, standard Qwen with impact reaches 0.446 macro-F1, 0.460 critical recall, and 0.383 worst-event macro-F1. Standard Qwen with BADGE reaches 0.468 macro-F1, 0.391 critical recall, and 0.374 worst-event macro-F1. Harrier with standard adaptation and BADGE has the highest 500-label worst-event F1 among these configurations (0.384), while frozen EmbeddingGemma with entropy has the highest 500-label critical recall (0.587). These are distinct Pareto points, not a single ranking.

6.2 Unseen-Event Robustness

Table 3 applies the same comparison to three unseen HumAID events. Harrier transfers the critical-recall benefit: impact exceeds random by 0.0571 AULC and also exceeds entropy by 0.0455 [0.0175, 0.0771] and Core-set by 0.0699 [0.0128,

Backbone	Δ Critical-recall AULC	Δ Macro-F1 AULC	Δ Worst-event F1 AULC
Qwen	+0.0885 [0.0585, 0.1177]	+0.0261 [0.0078, 0.0406]	+0.0324 [0.0011, 0.0552]
EmbeddingGemma	+0.0870 [0.0401, 0.1296]	+0.0189 [0.0048, 0.0332]	+0.0181 [−0.0047, 0.0341]
Harrier	+0.0575 [0.0210, 0.0897]	+0.0156 [−0.0022, 0.0368]	+0.0207 [0.0053, 0.0419]

Table 2: Untouched CrisisMMD confirmation: standard-adapter impact acquisition minus random sampling, paired by seed. Brackets give descriptive 95% bootstrap intervals; bold estimates have intervals excluding zero.

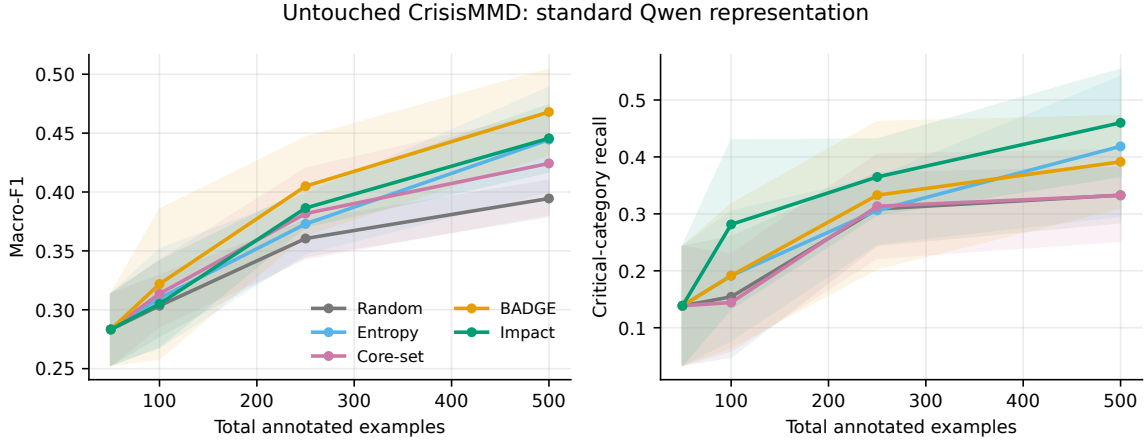


Figure 2: Mean CrisisMMD learning curves for standard-adapted Qwen; shading is one standard deviation across five paired seeds. All strategies share the same 50-label initialization within a seed. Impact acquisition improves critical recall while BADGE achieves the highest final macro-F1.

0.1270]. Its macro-F1 and worst-event F1 differences from random are near zero. EmbeddingGemma is directionally positive in critical recall but uncertain. Qwen does not transfer its CrisisMMD advantage.

The strongest held-out configurations are often frozen: frozen Qwen with entropy leads critical-recall AULC (0.515), whereas frozen EmbeddingGemma with random sampling leads macro-F1 (0.425) and worst-event macro-F1 AULC (0.423). This interaction is consistent with evidence that AL rankings depend strongly on the representation and pipeline (Ghose and Nguyen, 2024). It also rules out recommending impact acquisition as an “always-on” policy.

6.3 What Changes in the Labeled Set?

Figure 3 examines the 500-label CrisisMMD set. Random sampling selects 3.36% critical examples. With standard adaptation, impact acquisition raises this to 26.76% for Qwen, 34.16% for EmbeddingGemma, and 27.72% for Harrier. Qwen’s minimum number of selected examples per event rises from 16.2 to 39.6, while JSD from a uniform event allocation falls from 0.0921 to 0.0263. The corresponding JSD values are 0.0191 for EmbeddingGemma and 0.0256 for Harrier.

This is the intended mechanism: the policy constructs a labeled set that is more critical and event-balanced. It also deliberately departs from the naturally imbalanced pool. That intervention can reduce performance on common classes, explaining why critical recall and macro-F1 do not always move together.

Calibration effects are mixed. On CrisisMMD, impact lowers the binary critical Brier score at 500 labels for all three standard-adapted backbones (e.g., 0.0265 to 0.0145 for Qwen), but expected calibration error changes little and is not uniformly better. On held-out HumAID, critical Brier improves slightly for Harrier and EmbeddingGemma but worsens slightly for Qwen. We therefore make no general calibration claim.

6.4 Representation Adaptation Is Not Uniformly Helpful

Impact weighting inside the residual adapter is not consistently stronger than the standard adapter and is sometimes worse. The failure persists across acquisition policies and datasets. Even standard adaptation is not universally beneficial: frozen configurations lead several held-out outcomes. The positive headline result is consequently about co-designing acquisition with a con-

Backbone	Δ Critical-recall AULC	Δ Macro-F1 AULC	Δ Worst-event F1 AULC
Harrier	+0.0571 [0.0123, 0.0933]	+0.0048 [−0.0075, 0.0169]	+0.0075 [−0.0061, 0.0247]
EmbeddingGemma	+0.0424 [−0.0252, 0.1079]	−0.0097 [−0.0263, 0.0063]	−0.0062 [−0.0220, 0.0095]
Qwen	−0.0141 [−0.0518, 0.0313]	−0.0025 [−0.0257, 0.0283]	−0.0027 [−0.0315, 0.0272]

Table 3: Event-held-out HumAID: standard-adaptor impact acquisition minus random sampling. Harrier’s critical-recall gain transfers; Qwen’s does not.

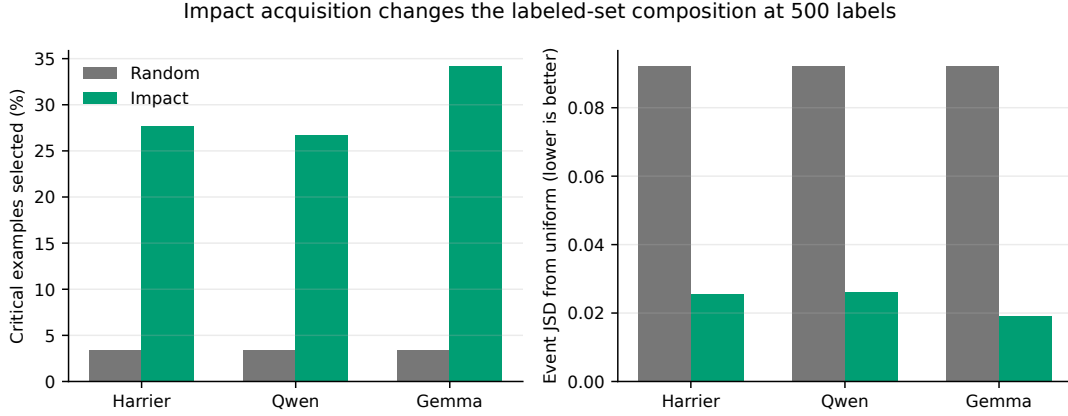


Figure 3: Composition of the standard-representation CrisisMMD labeled set at 500 labels. Impact acquisition selects substantially more critical examples and is closer to uniform event allocation (lower JSD).

servative representation, not about an intrinsically superior impact-weighted embedding objective.

The acquisition result itself changes by representation. For example, standard Qwen benefits strongly on CrisisMMD but not on unseen HumAID events, while standard Harrier transfers its critical-recall gain. Averaging across backbones would conceal this operationally relevant boundary. A deployment should validate the acquisition–representation pair on prior or simulated events rather than select the backbone and acquisition rule independently.

6.5 Resource and Compression Results

Table 4 reports measured extraction cost for the 11,584-item CrisisMMD training split. All models use substantially less than the 12 GB MIG allocation; EmbeddingGemma has the lowest peak allocation. The extraction times are similar under the shared batching configuration.

We truncate Qwen and EmbeddingGemma to 128, 256, and 512 prefix dimensions and renormalize. Under frozen random sampling, 512-dimensional Qwen halves storage and remains within 0.009 AULC of 1,024 dimensions on macro-F1, critical recall, and worst-event F1 across both test settings. EmbeddingGemma at 512 dimensions reduces storage by one third and

Backbone	Dim.	Seconds	Peak GiB
Harrier 0.6B	1,024	76.3	1.18
Qwen 0.6B	1,024	77.2	1.36
EmbeddingGemma 0.3B	768	73.4	0.67

Table 4: Measured embedding extraction on CrisisMMD train using one H100 MIG slice, batch size 16, and maximum length 192.

remains within 0.014. Even 128-dimensional Qwen, an 87.5% reduction, remains within 0.030 on all three outcomes across both datasets.

Compression is policy-dependent. With impact acquisition, 128-dimensional Qwen loses 0.053 critical-recall AULC on CrisisMMD relative to full dimension; the selector depends on probabilities and neighborhood geometry that a generic random linear probe does not test. These results support 512 dimensions for frozen random sampling in these testbeds; informed acquisition policies require separate compression checks.

7 Discussion

Impact is a property of the selection policy, not only the model. The same classifier family can receive qualitatively different supervision depending on which messages are annotated. On CrisisMMD, impact acquisition increases the labeled

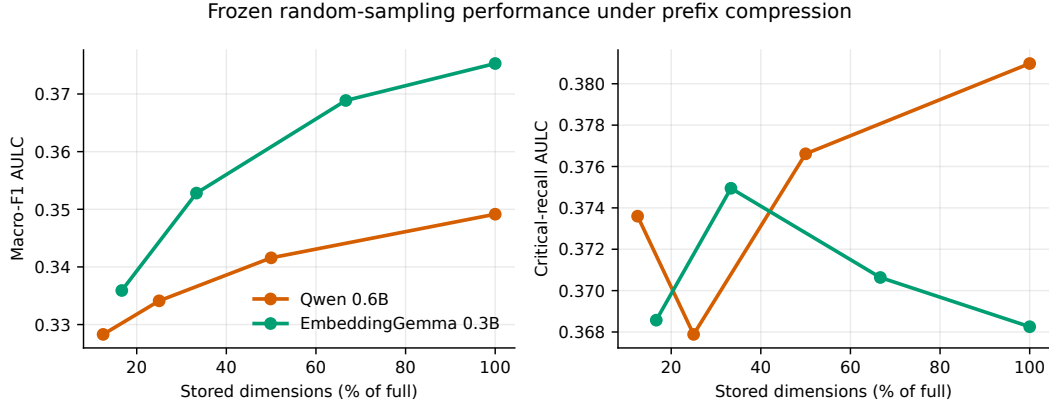


Figure 4: CrisisMMD AULC under prefix compression for frozen random sampling. The x-axis is stored dimensions as a percentage of the model’s full output. Critical recall is comparatively stable even when macro-F1 declines.

critical fraction by roughly eight to ten times and improves event balance. These composition metrics describe what the policy changes; they do not isolate the effects of individual score components or establish downstream humanitarian benefit.

There is no universally best operating point.

BADGE is a strong competitor for macro-F1; entropy can maximize critical recall with frozen embeddings; and impact acquisition provides a more balanced point for particular representation–dataset pairs. Presenting only a single metric or only the best backbone would overstate reliability. In practice, a response organization should specify whether missing critical cases, average category quality, or worst-event robustness is the binding constraint, then select from the observed Pareto frontier.

Representation validation must include the acquisition loop.

Embedding benchmarks and final linear-probe accuracy are insufficient for AL. Small geometric or calibration changes can alter which examples are selected, which changes every later training set. Our compression and held-out-event results both demonstrate this feedback. The relevant unit of validation is the full representation–acquisition pipeline.

Humanitarian use requires governance.

The critical taxonomy used here is a transparent task proxy. A real deployment should define priorities with affected communities and response practitioners, permit priorities to change by phase and context, audit false negatives, and retain human control over consequential decisions. Event balancing

is not a substitute for geographic, linguistic, or demographic equity.

8 Conclusion

We studied who receives scarce labels in humanitarian text classification. Impact-aware acquisition can substantially improve critical-category recall and event balance, and on untouched CrisisMMD it improves Qwen’s critical, aggregate, and worst-event AULC over random sampling. The benefit is not universal: strong baselines remain competitive, impact-weighted representation adaptation is inconsistent, and Qwen’s gain does not transfer to unseen HumAID events. These findings favor a representation-aware, Pareto-based deployment strategy over an always-on acquisition rule. Efficient open embeddings and prefix compression make the workflow feasible on a single MIG slice, but humanitarian utility must ultimately be validated with the people and organizations who define and bear the consequences of priority.

Limitations

First, CrisisMMD has only six missing/found examples in its test split. Its two-class critical macro recall is therefore sensitive to a handful of cases; the larger HumAID critical categories are an important counterweight. Second, our priority labels and event-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. Third, the method assumes event metadata is available. A deployment without event identifiers would need inferred geographic,

temporal, or semantic strata. Fourth, all annotation costs are treated as equal and we do not model annotator disagreement, expertise, or trauma exposure. Fifth, the datasets contain English social-media messages from a limited time period; results may not transfer to other languages, channels, disaster types, or humanitarian taxonomies. Sixth, our Core-set and BADGE baselines are efficient embedding-level approximations with a candidate cap. Seventh, five paired seeds permit transparent within-seed comparisons but do not justify strong asymptotic significance claims. Finally, the acquisition coefficients are fixed heuristics. Component and sensitivity ablations are valuable follow-up work, particularly before deployment.

Ethical and societal considerations

The study uses previously released crisis datasets and does not redistribute message text in the paper. 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 is intended to prioritize examples for human annotation, not to autonomously determine aid eligibility or operational response.

An impact-aware policy creates both benefits and risks. It can direct scarce annotation effort toward rare urgent categories, but an incorrect criticality model can repeatedly select sensational or ambiguous content. Event balancing can reduce dominance by large events, but equal event allocation may be inappropriate when severity and population exposure differ. Any deployment should log selection scores, monitor per-group and per-event 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 positive-impact claim is deliberately narrow: the method changes which examples receive labels and can improve benchmark recall on predefined critical classes. It does not establish a positive societal outcome. Validating that link requires participatory priority definition, practitioner evaluation, and prospective deployment studies.

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.
- 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.
- 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.

Microsoft. 2026. [Harrier-OSS-v1 model card](#). Hugging Face model documentation.

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.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panayam, 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: Debiased 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 backbones; frozen, standard, and impact-weighted representations; five acquisition strategies; five seeds; budgets 50, 100, 250, and 500; primary outcomes; candidate cap; and paired-bootstrap comparisons. We also fixed the interpretation rule that no method would be called dominant unless critical recall and macro-F1 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
Event-balance coefficient	0.10
Logistic regression C	1.0
Bootstrap resamples	10,000

Table 5: Fixed experimental hyperparameters.

C Additional Reproducibility Details

Each result JSON records the dataset, backbone, full or truncated embedding dimension, representation, strategy, 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. 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.