Cross-disaster Domain Adaptation Using Co-training Variants

Khushboo Gupta
University of Illinois at Chicago
kgupta27@uic.edu

Jacob Ativo
California State University, East Bay
jativo@horizon.csueastbay.edu

Anh Tran
Independent Researcher
anhtranst@gmail.com

Hongmin Li
California State University, East Bay
hongmin.li@csueastbay.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

Cornelia Caragea
University of Illinois at Chicago
cornelia@uic.edu

ABSTRACT

Automated classification of crisis-related social media posts is widely used to support humanitarian response; however, models trained on historical disasters often degrade when applied to new events due to cross-disaster domain shift. In emerging crises, labeled data is scarce while large volumes of unlabeled content accumulate rapidly, making effective domain adaptation critical for reliable deployment. In this work, we investigate semi-supervised domain adaptation for cross-disaster tweet classification under temporally realistic transfer settings, where each target event occurs strictly later than its source event. We evaluate adaptation performance across multiple humanitarian disasters under low-data regimes (5–50 labeled examples per class), distinguishing between within-disaster and cross-disaster transfer.

We compare fully supervised fine-tuning, self-training, and unsupervised domain adaptation (UDA) against a structured co-training framework that leverages dual-view source–target supervision and cross-view pseudo-label exchange. We further study a family of controlled design variations that modify individual components—such as pseudo-label selection and mixup regularization—to analyze their impact on cross-event generalization and calibration. Results show these co-training variants consistently outperform UDA alone in low-resource settings, while different pseudo-label utilization strategies exhibit distinct trade-offs across disaster types and label budgets. By providing a temporally grounded benchmark and a structured analysis of adaptation mechanisms, this work contributes empirical guidance for designing more robust cross-disaster classification systems for crisis informatics.

Keywords

Domain adaptation, co-training, crisis informatics, social media classification

INTRODUCTION

Tools for automating the classification of disaster-related social media data have become indispensable for humanitarian response, enabling timely identification of relevant information from streams of user-generated text. However, most supervised language models rely on labeled data, which is often difficult to obtain during crisis situations. Moreover, models trained on past events frequently fail to generalize to newly emerging disasters due to substantial differences in language, damage patterns, and temporal and geographic conditions. Early work on crisis tweet classification (Imran, Mitra, et al. 2016) shows that re-purposing labeled data from prior disasters can improve performance when source and target events are similar, but distributional mismatch remains a central challenge for real-world deployment across heterogeneous crises.

WiP Paper – Social Media & Crisis Communication: narratives, signals, and sentiments
Proceedings of the 23rd ISCRAM Conference – The Hague, the Netherlands June 2026
Caroline Rizza, Apoorva Chauhan, Amy Matser, Joyce Kox, Willem Treurniet and Jeroen Wolbers eds.

Domain shift—i.e., differences in feature distributions between a historical (source) event and a current (target) event—creates major challenges for reliable transfer learning in crisis informatics. When the shift is large, naive transfer can reduce classification accuracy and induce overconfident errors, weakening trust in automated decision-making in high-stakes settings. In the broader machine learning literature, unsupervised domain adaptation (UDA) (Ganin and Lempitsky 2015a; Chhabra et al. 2024) seeks to narrow this gap by aligning source and target representations using adversarial or discrepancy-based objectives (Luo et al. 2022; Zeng et al. 2024), without requiring target labels. However, UDA alone typically addresses only global distribution mismatch and provides no direct way to exploit even limited target supervision when it is available.

In contrast, semi-supervised domain adaptation (SSDA) and pseudo-labeling strategies (Yu and Lin 2023; Hosseini and C. Caragea 2023) aim to leverage a small set of labeled target examples alongside large pools of unlabeled target data to improve representation learning and model calibration. Pseudo-labeling methods (Lee 2013; Arazo et al. 2020) iteratively assign labels to unlabeled instances, effectively expanding the supervision signal, but they are sensitive to label noise and confirmation bias if errors are reinforced over time (H. Li, Wu, et al. 2021; Arazo et al. 2020). These challenges can be amplified in crisis informatics, where linguistic variation, rapidly evolving contexts, and class imbalance are common. As a result, approaches developed primarily for vision datasets or generic domain benchmarks (e.g., Office-31, DomainNet) may not transfer to disaster-related text (Ben-David et al. 2010; Wang and Deng 2018), motivating adaptation methods that explicitly account for uncertainty and event-specific shift.

Further, model calibration (Guo et al. 2017; Minderer et al. 2021), which quantifies the agreement between predicted posterior probabilities and empirical correctness rates, ensuring probabilistic reliability beyond accuracy alone, is particularly important in emergency response settings, where overconfident predictions can mislead responders. In crisis informatics (Imran, Mitra, et al. 2016; Olteanu et al. 2014), decision-makers often rely on probability scores to triage information, allocate resources, and prioritize follow-up actions, so poorly calibrated outputs can translate into real operational risk. Recent advances in UDA calibration show that standard adaptation techniques can still yield miscalibrated models under domain shift (Ovadia et al. 2019; Kumar et al. 2019), motivating methods that explicitly manage predictive uncertainty.

To address these challenges, we pursue a structured study of Deep Co-Training with Task Decomposition (DeCoTa) (L. Yang et al. 2021) as a semi-supervised domain adaptation framework for cross-disaster tweet classification. DeCoTa differs from single-model or single-view adaptation methods by jointly training multiple classifier views that exchange pseudo-labels and are regularized through cross-view consistency objectives, drawing on modern deep co-training and multi-view learning principles (Blum and Mitchell 1998; Zhou and M. Li 2005). This multi-view training paradigm mitigates confirmation bias and promotes more reliable exploitation of unlabeled target data. In the crisis-domain setting, such structure is particularly advantageous: cross-view agreement discourages the propagation of low-confidence pseudo-labels and provides an implicit uncertainty filter, thereby stabilizing learning under event-specific distribution shift. Building on this foundation, we introduce a family of component-level design variations that systematically modulate pseudo-label selection criteria, mixup-based augmentation (H. Zhang et al. 2018), confidence weighting, and optional language-model guidance. Rather than conventional ablations, these constitute targeted design interventions that isolate the functional contribution of each mechanism to cross-event generalization, with particular emphasis on model efficiency and probabilistic calibration under domain shift.

We evaluate our methods on a temporally ordered cross-disaster adaptation setup spanning multiple historical disasters, where training is performed on earlier events and adaptation targets strictly later events, reflecting realistic deployment constraints in crisis response. Experiments are conducted in low-label target regimes (5 to 50 labeled examples per class from the target domain), simulating early-stage disaster scenarios in which annotation resources are limited (Alam et al. 2021). Performance is assessed using Macro-F1 to assess class-balanced predictive effectiveness and Expected Calibration Error (ECE) to quantify probabilistic reliability under distribution shift. This dual evaluation framework captures not only classification performance but also the operational reliability of model confidence estimates in high-stakes settings. Our contributions are threefold:

1. **Temporally grounded cross-disaster evaluation:** We establish a temporally ordered cross-disaster adaptation protocol that distinguishes within-event (same-disaster) transfer from cross-event (out-of-disaster) generalization, enforcing strict train-on-past, adapt-to-future constraints to reflect realistic crisis deployment settings.

2. **Structured analysis of DeCoTa-based adaptation mechanisms:** We develop a family of DeCoTa-based variants that systematically vary pseudo-label selection, mixup-based augmentation strategies, confidence reweighting, and LLM-guided refinement, enabling controlled analysis of their contributions to cross-event generalization under domain shift.

3. **Empirical analysis of performance–reliability trade-offs:** We provide a comprehensive empirical comparison of domain adaptation, semi-supervised learning, and structured co-training approaches, quantifying trade-offs between predictive performance and probabilistic calibration across varying target label budgets and disaster types.

RELATED WORK

Automated classification of crisis-related social media posts has been widely studied as a means of extracting actionable information during disasters. Early systems such as CrisisLex and AIDR (Olteanu et al. 2014; Imran, Castillo, Lucas, et al. 2014) demonstrated that supervised machine learning models can effectively categorize tweets into humanitarian classes including infrastructure damage, requests for aid, and situational reports, but performance degrades under severe event disparity, motivating domain adaptation for crisis informatics. For example, Imran, Mitra, et al. (2016) utilized cross-language domain adaptation for classifying crisis messages and found that adaptation effectiveness varies significantly with domain similarity. H. Li, D. Caragea, C. Caragea, and Herndon (2017) showed that domain adaptation methods improve tweet classification quality in the context of disaster response. Recent work continues to explore AI-enhanced frameworks for tweet classification (Karimiziarani et al. 2025), indicating the ongoing challenge of efficient generalization across heterogeneous disaster events.

In the broader field, unsupervised domain adaptation (UDA) addresses distribution gaps by aligning source and target feature representations without using labeled target data. While classical adversarial and discrepancy-based techniques can reduce domain mismatch (Saito et al. 2018; Y. Zhang et al. 2019), they often struggle under class imbalance and semantic shifts typical in real-world crises. Domain adaptation methods, specifically for disaster tweet classification have also been proposed using reconstruction networks and adversarial alignment (X. Li and D. Caragea 2020). Semi-supervised Domain Adaptation (SSDA), where a small number of labeled target examples are available, have been formally studied and shown to outperform UDA in leveraging limited target supervision for better feature alignment (Y. Kim and C. Kim 2021). However, performance remains sensitive to label scarcity and class imbalance—conditions that are common in emerging disaster events.

Semi-supervised learning (SSL) aims to improve model performance by leveraging unlabeled data in addition to labeled samples. One of the simplest and most widely adopted SSL strategies is pseudo-labeling, where high-confidence predictions on unlabeled data are iteratively added to the training set (Lee 2013). More recent methods incorporate consistency regularization, encouraging models to produce stable predictions under input perturbations (Laine and Aila 2016; Sohn et al. 2020). While SSL has achieved strong results in low-label regimes, pseudo-label noise remains a major challenge. Arazo et al. (2020) show that confirmation bias can significantly degrade performance when pseudo-labels are incorrect. Under domain shift, this issue is amplified because early model predictions are influenced by source-domain features that may not transfer cleanly to the target domain. In crisis informatics contexts, where linguistic variation and rare humanitarian categories are common, pseudo-label errors can disproportionately affect minority classes and degrade calibration (Imran, Castillo, Diaz, et al. 2015; Guo et al. 2017; Arazo et al. 2020).

Co-training was originally proposed as a semi-supervised framework that exploits multiple conditionally independent and sufficient views of the data to iteratively assign pseudo-labels to unlabeled samples (Blum and Mitchell 1998). Deep co-training extends this idea to neural networks by jointly training multiple classifiers with cross-view agreement objectives while encouraging diversity through architectural or stochastic variation (Qiao et al. 2018; Zhou and M. Li 2005). By requiring consensus across views, multi-view learning mitigates confirmation bias relative to single-model self-training, as erroneous pseudo-labels must be supported by more than one predictor. Recent work has begun integrating co-training with domain adaptation objectives, demonstrating improved robustness under distribution shift compared to single-model pseudo-labeling (Saito et al. 2018). However, a systematic examination of co-training strategies in temporally ordered cross-disaster text classification remains limited.

Mixup regularization (H. Zhang et al. 2018) improves generalization by interpolating input representations and labels, encouraging smoother decision boundaries. Variants of mixup have been adapted for semi-supervised learning (Berthelot et al. 2019) and domain adaptation, including inter-domain mixup strategies that blend source and target samples to improve alignment. Confidence-aware learning further refines pseudo-label utilization by weighting contributions according to model certainty (Xie et al. 2020). Such strategies aim to mitigate the impact of noisy pseudo-labels while still exploiting unlabeled data. In cross-disaster contexts, confidence-aware mechanisms are particularly relevant due to the presence of ambiguous tweets and rare categories. However, the interaction between mixup-based smoothing, pseudo-label confidence, and cross-event shift remains underexplored in humanitarian classification tasks.

Large language models (LLMs) have recently demonstrated strong zero-shot and few-shot capabilities for text classification tasks. Instruction-tuned LLMs can generalize across domains by leveraging large-scale pretraining on

diverse corpora (Brown et al. 2020; Touvron et al. 2023). In crisis informatics, LLMs have been explored for disaster tweet classification (Yin et al. 2024) and summarization, showing potential for rapid deployment in low-label settings. Nevertheless, LLM predictions may exhibit instability or systematic bias under domain-specific constraints, and their calibration properties remain inconsistent across tasks (Xu et al. 2025; Lei et al. 2025). Integrating LLM outputs into structured adaptation pipelines therefore requires careful evaluation, particularly when used to guide pseudo-label refinement under distribution shift.

Building on these prior works, we investigate cross-disaster adaptation through a structured analysis of DeCoTa-based mechanisms under temporally realistic transfer settings. In contrast to prior approaches that emphasize standalone UDA, SSL, or isolated pseudo-labeling strategies, we systematically examine how co-training dynamics, mixup variants, confidence reweighting, and optional LLM-guided refinement affect cross-event generalization and probabilistic calibration under limited target supervision.

METHODS

We investigate domain adaptation under limited target supervision by comparing supervised fine-tuning, self-training (ST), UDA, and a structured family of DeCoTa-based variants (L. Yang et al. 2021). All methods operate under a unified setting consisting of labeled source data D_L^S, a small labeled target subset D_L^T, and a larger unlabeled target pool D_U^T. To ensure fair architectural comparison, all models employ *BERTweet* (Nguyen et al. 2020) as the shared encoder backbone. Detailed descriptions of each method are provided in the following subsections.

Baselines

We first evaluate three baseline approaches—Supervised Fine-Tuning, Self-Training (ST), and Unsupervised Domain Adaptation (UDA)—to provide controlled reference points for adaptation under limited target supervision.

Zero-shot LLM (Qwen3-14B)

The zero-shot LLM baseline uses Qwen3-14B (A. Yang et al. 2025) to classify tweets into predefined humanitarian categories without any task-specific training on target data. Predictions are generated using a structured prompt with a constrained output format. Unlike other methods, this approach does not use labeled or unlabeled target data for adaptation. Qwen3-14B (14B parameters) is substantially larger than the BERTweet backbone used in our framework, and serves as a strong general-purpose reference. This baseline reflects performance under pretrained language knowledge alone and provides a comparison to methods that leverage target data or distill LLM signals into smaller, task-specific models.

Supervised Fine-Tuning

Supervised fine-tuning uses only the limited labeled target subset D_L^T. The BERTweet encoder is trained using standard cross-entropy loss over these labeled target instances, without leveraging source data or unlabeled target examples. This baseline represents adaptation under strictly low-resource supervision and serves as a lower bound for cross-disaster transfer. Because it relies solely on scarce target annotations, performance is highly sensitive to label scarcity and class imbalance.

Self-Training (ST)

Self-Training (H. Li, D. Caragea, and C. Caragea 2021) extends supervised fine-tuning by incorporating unlabeled target data D_U^T through iterative pseudo-labeling. After initial training on D_L^T, the model generates pseudo-labels for unlabeled target instances. Predictions exceeding a confidence threshold are retained and treated as additional labeled data in subsequent optimization. This refinement process is repeated over multiple iterations. By expanding supervision within the target domain, ST can improve performance over purely supervised learning. However, because pseudo-labels are generated by a single model view, errors may be reinforced across iterations under domain shift, making the approach susceptible to confirmation bias.

Unsupervised Domain Adaptation (UDA)

Unsupervised Domain Adaptation leverages labeled source data D_L^S together with unlabeled target data D_U^T, without using labeled target supervision (Ganin and Lempitsky 2015b). The model is trained using supervised loss on the source domain and an additional regularization term over unlabeled target instances to encourage prediction consistency or feature-level alignment between domains. Unlike Self-Training, UDA transfers knowledge from prior disasters through source supervision but does not incorporate explicit target labels. Its effectiveness therefore depends on the magnitude of distribution shift between source and target events, and alignment alone may be insufficient when domain gaps are substantial.

DeCoTa

DeCoTa (L. Yang et al. 2021) integrates source supervision, target supervision, and unlabeled target adaptation within an asymmetric co-training framework. Two classifier views are maintained: f_1, supervised on labeled source data D_L^S, and f_2, supervised on labeled target data D_L^T.

$$\mathcal{L}_{sup}^{(1)} = \mathbb{E}_{(x,y)\sim D_L^S} \ell(f_1(x), y), \quad \mathcal{L}_{sup}^{(2)} = \mathbb{E}_{(x,y)\sim D_L^T} \ell(f_2(x), y). \tag{1}$$

For $x \in D_U^T$, each classifier produces pseudo-labels when confidence exceeds threshold τ. These pseudo-labels are exchanged across views, such that pseudo-labels generated by f_1 supervise f_2, and vice versa. To regularize pseudo-label noise, DeCoTa applies view-consistent mixup. Pseudo-labeled target samples used to update f_1 are mixed with labeled source data, while those used to update f_2 are mixed with labeled target data:

$$\tilde{x} = \lambda x_a + (1-\lambda) x_b, \quad \tilde{y} = \lambda y_a + (1-\lambda) y_b, \tag{2}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$. The final objective combines supervised loss and mixup-regularized pseudo-label loss across both views, and predictions are ensembled at inference time.

To isolate the contribution of individual design choices, we introduce controlled variants that modify specific components of the base DeCoTa framework.

Low-confidence mixup

Unlike the base formulation, which retains pseudo-labels exceeding confidence threshold τ, this variant instead leverages pseudo-labels with confidence *below* the threshold. Formally, the pseudo-labeled set becomes

$$\tilde{D}_U^T = \{(x, \hat{y}) \mid x \in D_U^T,\ \max f(x) < \tau\}. \tag{3}$$

By incorporating lower-confidence predictions, this variant emphasizes uncertain target instances that lie closer to the decision boundary. The objective remains structurally identical, but supervision is shifted toward harder examples. This encourages boundary refinement and can improve calibration behavior, though it may slow convergence and reduce peak performance in extremely low-label regimes.

Cross-View mixup

The cross-view mixup variant modifies the pairing strategy in the mixup regularization step. In the base formulation, pseudo-labeled samples are mixed with labeled data consistent with each classifier's supervisory anchor (source for f_1, target for f_2). Here, the pairing is reversed: pseudo-labels for the source-supervised classifier are mixed with labeled target data, and vice versa. This enforces stronger cross-domain interpolation during pseudo-label refinement, encouraging domain alignment but potentially reducing domain anchoring when supervision is extremely limited.

Weighted by confidence mixup

Instead of binary filtering, this variant scales pseudo-label loss according to prediction confidence:

$$\mathcal{L}_{pseudo} = \mathbb{E}\big[w(x)\, \ell(f(x), \hat{y})\big], \quad w(x) = \max f(x). \tag{4}$$

High-confidence samples therefore contribute more strongly during optimization. This typically accelerates adaptation and improves peak F1 performance, but may amplify early overconfidence and increase ECE under severe domain shift.

LLM-start

In the LLM-start variant, we use Qwen3-14B (A. Yang et al. 2025) to provide pseudo-labels for unlabeled target instances during the initial adaptation phase, and then transition to the original DeCoTa procedure. Let $\hat{y}^{LLM}(x)$ denote the LLM-generated pseudo-label for $x \in D_U^T$, and let $\hat{y}^{Dec}(x)$ denote the pseudo-label produced by the DeCoTa classifier view used for cross-view exchange. We apply a three-stage schedule over refinement iterations: (i) in the first iteration, pseudo-label supervision on D_U^T uses only $\hat{y}^{LLM}$; (ii) in the second iteration, pseudo-label supervision uses a 50/50 mixture of LLM-based and classifier-based pseudo-labels; and (iii) from the third iteration onward, we revert to standard DeCoTa pseudo-label exchange exclusively. This warm-start provides external guidance when target predictions are least reliable, while preserving DeCoTa's self-training dynamics in later stages.

Source	SD Target	OOD Targets
Canada Wildfires 2016 (8)	California Wildfires 2018 (10)	Kerala Floods 2018 (9); Hurricane Dorian 2019 (9)
Hurricane Harvey 2017 (9)	Hurricane Dorian 2019 (9)	Kerala Floods 2018 (9); California Wildfires 2018 (10)
Hurricane Irma 2017 (9)	Cyclone Idai 2019 (10)	Kerala Floods 2018 (9); California Wildfires 2018 (10)
Kaikoura Earthquake 2016 (9)	–	Hurricane Florence 2018 (9); California Wildfires 2018 (10); Kerala Floods 2018 (9)
Kerala Floods 2018 (9)	–	Hurricane Dorian 2019 (9); California Wildfires 2018 (10)

Table 1. Domain adaptation configuration. Each row lists a source event and its corresponding same-disaster (SD) and out-of-disaster (OOD) target events. Numbers in parentheses indicate the number of humanitarian classes per dataset.

EXPERIMENTAL SETUP

This section details the experimental setup, including dataset configuration, evaluation metrics, and training hyperparameters.

Dataset

For this work, we utilize selected events from the HumAID dataset (Alam et al. 2021) to construct a cross-disaster domain adaptation benchmark. Rather than using all events jointly (Gupta et al. 2025), we define structured source–target splits where five disaster events serve as sources and are they are evaluated against same-disaster (SD) or out-of-disaster(OOD) target disaster events. The number of classes per event ranges from 8 to 10. Specifically, we consider Canada Wildfires, Hurricane Harvey, Hurricane Irma, Kaikoura Earthquake, and Kerala Floods as source events. Each source is paired with one or more SD and OOD target disasters, including California Wildfires, Hurricane Dorian, Cyclone Idai, Hurricane Florence, and others, as summarized in Table 1.

Cross-Disaster Domain Adaptation Setup

We evaluate cross-disaster domain adaptation for humanitarian tweet classification using temporally ordered transfer pairs $S \rightarrow T$, where a fully labeled source event S is used to adapt to a later target event T. Temporal ordering prevents information leakage and reflects realistic deployment from past to emerging crises. For each source event, we define two target settings. In the SD (same-disaster) setting, the target is a later event of a similar disaster type, capturing within-category temporal or geographic variation. In the OOD (out-of-disaster) setting, the target represents a different disaster type, modeling cross-hazard distribution shift.

Target training data is partitioned into a small labeled subset D_L^T and a larger unlabeled pool D_U^T. We construct k-shot labeled subsets with $k \in \{5, 10, 25, 50\}$ samples per class, repeated across three random seeds. Remaining target training instances are treated as unlabeled. Validation and test splits follow the original dataset partitions. We assume a shared humanitarian label space across events while allowing each target to contain only a subset of labels. Tweets are used in raw form with minimal normalization to preserve real-world linguistic variability.

RESULTS AND DISCUSSION

We evaluate supervised fine tuning, ST, UDA, and DeCoTa-based variants across multiple source–target disaster pairs and label regimes (5, 10, 25, and 50 labeled samples per class). When results are averaged across all source–target pairs (Table 2), the zero-shot LLM achieves the highest F1 at 5 and 10 labels per class, highlighting strong generalization in extremely low-label settings. However, this advantage is less consistent when results are examined per source–target pair (see Tables T1-T5 in the Appendix), where the LLM baseline is only consistently strongest at 5 labels per class. Across all experiments, macro-F1 increases consistently with additional supervision, confirming that even modest target annotations substantially improve cross-domain generalization. However, the magnitude and stability of improvement differ across methods.

Supervised learning improves steadily but is consistently outperformed by semi-supervised and domain-adaptive approaches. Self-Training provides competitive performance at moderate and high label regimes, while UDA offers stable gains but is generally surpassed by DeCoTa variants at 10 labels and above. These results suggest that structured pseudo-label refinement yields more effective transfer than UDA alone. While zero-shot Qwen3-14B achieves stronger average F1 at 5 and 10 labels per class, this advantage is limited to extremely low-label regimes. Within the DeCoTa family, distinct tradeoffs emerge. Weighted-confidence often achieves the highest peak F1 at 25–50 labels per class and consistently outperforms the LLM baseline in these regimes but exhibits calibration instability in low-label settings. In contrast, the low-confidence variant yields more conservative performance

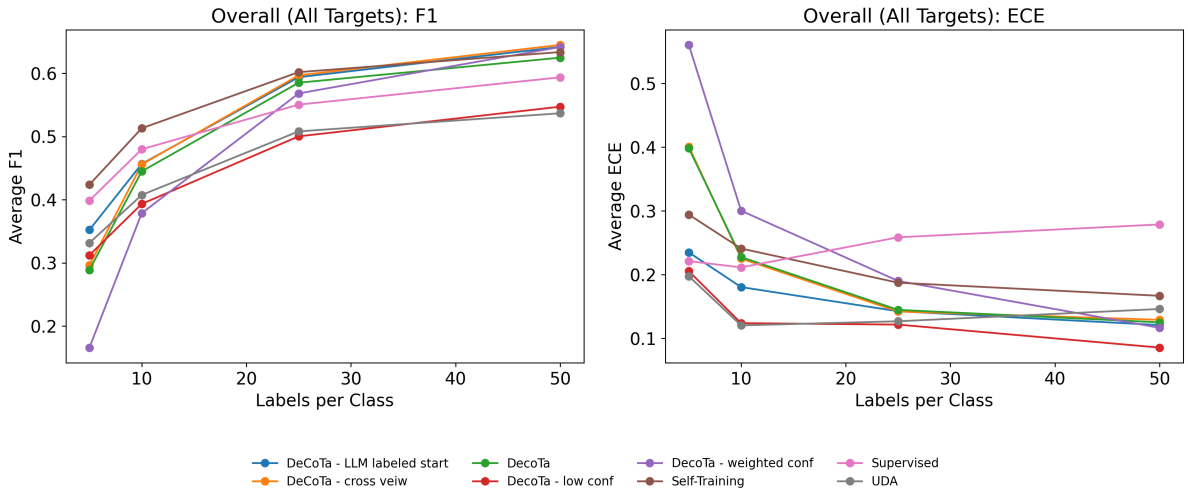

Figure 1. Overall F1 and ECE summary across target events.

with comparatively lower ECE. Cross-view provides the most balanced behavior, achieving near-peak F1 while maintaining more stable calibration across label regimes. The LLM-initialized variant offers clear benefits at 5 labels per class, indicating that language-model-informed initialization is particularly useful in extremely low-supervision scenarios, though its relative advantage diminishes as labeled data increases.

Calibration analysis further reveals non-monotonic ECE behavior, especially at 5 labels per class. Aggressive confidence weighting tends to amplify early overconfidence, whereas cross-view and low-confidence strategies mitigate this effect. As supervision increases to 25 and 50 labels, calibration stabilizes across nearly all methods. Target-specific trends are consistent with these observations. DeCoTa variants dominate moderate-to-high label regimes across disasters, while early-label calibration volatility is most pronounced for weighted-confidence approaches. Label-wise analysis (Figure S2) shows that while overall performance improves with additional supervision, gains are uneven across classes, with certain categories remaining consistently difficult despite increased labeled data. In particular, classes such as *caution_and_advice*, *infrastructure_and_utility_damage*, and *requests_or_urgent_needs* exhibit persistently lower performance, reflecting higher semantic ambiguity and overlap with other categories. In contrast, categories such as *displaced_people_and_evacuations*, *rescue_volunteering_or_donation_effort*, and *sympathy_and_support* achieve relatively stronger performance, indicating clearer lexical patterns.

In summary, structured co-training mechanisms substantially improve cross-disaster adaptation over standard UDA baselines, while introducing tunable trade-offs between peak performance and calibration reliability. These trade-offs are particularly important in crisis informatics settings, where both predictive accuracy and well-calibrated uncertainty estimates are critical for downstream decision-making. Figure 1 presents overall adaptation performance averaged across all targets. Table 2 reports performance averaged across sources for each target and domain adaptation methods. Because these averages aggregate results over multiple source domains—including some with limited transferability to specific targets—the aggregated scores are comparatively lower than the strongest individual source–target pairings. Nevertheless, structured co-training consistently outperforms baseline approaches at the individual source–target level. Detailed per source–target results are provided in the appendix.

CONCLUSION AND FUTURE WORK

This study evaluated domain-adaptive and semi-supervised approaches for cross-disaster tweet classification under limited supervision. Across multiple source–target pairs and label regimes, DeCoTa-based variants consistently outperformed Qwen3-14B zero-shot, supervised learning, self-training, and UDA, particularly at moderate-to-high label counts. While all methods benefited from additional supervision, structured co-training mechanisms enabled more effective cross-domain transfer and revealed important performance–calibration tradeoffs. In particular, cross-view training provided the most balanced improvement, achieving strong predictive performance while maintaining comparatively stable calibration.

Our findings have direct implications for the deployment of automated social media classification systems in crisis response. In practice, models trained on historical disasters should not be applied to new events without adaptation, as cross-disaster domain shift can significantly degrade performance. Instead, practitioners should prioritize rapid

annotation of a small, representative subset of target data early in an event, as even a limited number of labeled samples can substantially improve model effectiveness when used with structured semi-supervised approaches. Unlabeled data, while abundant, must be leveraged carefully using mechanisms that control pseudo-label noise and uncertainty, rather than naive self-training. Additionally, model confidence scores should be interpreted cautiously, as overconfidence under domain shift may lead to misleading prioritization of information. These observations suggest that adaptive, human-in-the-loop systems—where models are continuously updated with new data and used to assist rather than replace analysts—are best suited for reliable deployment in dynamic crisis environments.

Future work will focus on leveraging the observed strengths of individual DeCoTa variants to design a unified, more adaptive algorithm. Our results reveal complementary behaviors across variants: weighted-confidence achieves strong peak performance, low-confidence improves calibration stability, cross-view provides balanced robustness, and LLM-initialized training benefits early-label regimes. A promising direction is to integrate these mechanisms within a dynamically controlled framework that adjusts confidence weighting, agreement constraints, and pseudo-label selection based on model readiness and uncertainty signals during training. Such an adaptive strategy could mitigate early overconfidence while preserving high-label performance gains, leading to a more stable and consistently superior cross-domain adaptation algorithm. Another important direction for future work is the inclusion of strong LLM baselines, such as few-shot and chain-of-thought prompting. While this work uses LLMs for initialization, evaluating them as standalone classifiers would enable a more complete comparison and inform how LLM-based and semi-supervised approaches can be combined.

In addition, while our temporally ordered cross-disaster evaluation protocol enforces realistic deployment constraints across events, extending this temporal structure within individual disasters remains an open direction. Disaster dynamics evolve rapidly over time, and modeling intra-event temporal shifts could provide a more fine-grained understanding of adaptation under continuously changing conditions. Incorporating such temporal progression into training and evaluation may further improve the realism and robustness of crisis informatics systems. However, the current dataset does not include explicit temporal annotations, limiting our ability to model or evaluate temporal dynamics.

ACKNOWLEDGMENT

This work is supported by a collaborative CAHSI-Google Institutional Research Program award.

REFERENCES

Alam, F., Qazi, U., Imran, M., and Ofli, F. (2021). "HumAID: Human-Annotated Disaster Incidents Data from Twitter". In: *15th International Conference on Web and Social Media (ICWSM)*.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2020). "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2010). "A Theory of Learning from Different Domains". In: *Machine Learning* 79.1-2, pp. 151–175.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). "MixMatch: A Holistic Approach to Semi-Supervised Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Blum, A. and Mitchell, T. (1998). "Combining labeled and unlabeled data with co-training". In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*. COLT' 98. Madison, Wisconsin, USA: Association for Computing Machinery, pp. 92–100.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). "Language models are few-shot learners". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc.

Chhabra, S., Venkateswara, H., and Li, B. (2024). "Domain Adaptation Using Pseudo Labels". In: *ArXiv* abs/2402.06809.

Ganin, Y. and Lempitsky, V. (July 2015a). "Unsupervised Domain Adaptation by Backpropagation". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 1180–1189.

Ganin, Y. and Lempitsky, V. (2015b). "Unsupervised domain adaptation by backpropagation". In: *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, pp. 1180–1189.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 1321–1330.

Gupta, K., Gautam, N., Sosea, T., Caragea, D., and Caragea, C. (May 2025). "Calibrated Semi-Supervised Models for Disaster Response based on Training Dynamics". In: *Proceedings of the International ISCRAM Conference*.

Hosseini, M. and Caragea, C. (July 2023). "Semi-Supervised Domain Adaptation for Emotion-Related Tasks". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 5402–5410.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (June 2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Comput. Surv.* 47.4.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: artificial intelligence for disaster response". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, pp. 159–162.

Imran, M., Mitra, P., and Srivastava, J. (2016). "Cross-language domain adaptation for classifying crisis-related short messages". English (US). In: *ISCRAM 2016 Conference Proceedings - 13th International Conference on Information Systems for Crisis Response and Management*. Ed. by P. Antunes, V. A. Banuls Silvera, J. Porto de Albuquerque, K. A. Moore, and A. H. Tapia. Proceedings of the International ISCRAM Conference. Information Systems for Crisis Response and Management, ISCRAM.

Karimiziarani, M., Foroumandi, E., and Moradkhani, H. (2025). "Harnessing Twitter (X) with AI-enhanced natural language processing for disaster management: Insights from California wildfire". In: *Environmental Modelling and Software* 192, p. 106545.

Kim, Y. and Kim, C. (Jan. 2021). " Semi-Supervised Domain Adaptation via Selective Pseudo Labeling and Progressive Self-Training ". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 1059–1066.

Kumar, A., Liang, P. S., and Ma, T. (2019). "Verified Uncertainty Calibration". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Laine, S. and Aila, T. (2016). "Temporal Ensembling for Semi-Supervised Learning". In: *CoRR* abs/1610.02242. arXiv: 1610.02242.

Lee, D.-H. (July 2013). "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks". In: *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.

Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q. R., and Li, J. (July 2025). "Harnessing Large Language Models for Disaster Management: A Survey". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 14528–14551.

Li, H., Wu, Z., Shrivastava, A., and Davis, L. S. (2021). "Rethinking Pseudo Labels for Semi-Supervised Object Detection". In: *AAAI Conference on Artificial Intelligence*.

Li, H., Caragea, D., and Caragea, C. (2021). "Combining self-training with deep learning for disaster tweet classification". In: *The 18th international conference on information systems for crisis response and management (ISCRAM 2021)*.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2017). "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26, pp. 16–27.

Li, X. and Caragea, D. (2020). "Domain Adaptation with Reconstruction for Disaster Tweet Classification". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China: Association for Computing Machinery, pp. 1561–1564.

Luo, Y.-W., Ren, C.-X., Dai, D.-Q., and Yan, H. (Mar. 2022). " Unsupervised Domain Adaptation via Discriminative Manifold Propagation ". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 44.03, pp. 1653–1669.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. (2021). "Revisiting the calibration of modern neural networks". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A. (Oct. 2020). "BERTweet: A pre-trained language model for English Tweets". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Online: Association for Computational Linguistics, pp. 9–14.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (May 2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 376–385.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. (2019). "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Qiao, S., Shen, W., Zhang, Z., Wang, B., and Yuille, A. (2018). "Deep Co-Training for Semi-Supervised Image Recognition". In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*. Munich, Germany: Springer-Verlag, pp. 142–159.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (June 2018). "Maximum Classifier Discrepancy for Unsupervised Domain Adaptation". In: pp. 3723–3732.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. (2020). "FixMatch: simplifying semi-supervised learning with consistency and confidence". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). "LLaMA: Open and Efficient Foundation Language Models". In: *ArXiv* abs/2302.13971.

Wang, M. and Deng, W. (Oct. 2018). "Deep visual domain adaptation: A survey". In: *Neurocomput.* 312.C, pp. 135–153.

Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2020). "Unsupervised data augmentation for consistency training". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 6256–6268.

Xu, F., Ma, J., Li, N., and Cheng, J. C. (2025). "Large language model applications in disaster management: An interdisciplinary review". In: *International Journal of Disaster Risk Reduction* 127, p. 105642.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). *Qwen3 Technical Report*. arXiv: 2505.09388 [cs.CL].

Yang, L., Wang, Y., Gao, M., Shrivastava, A., Weinberger, K. Q., Chao, W.-L., and Lim, S.-N. (2021). "Deep Co-Training with Task Decomposition for Semi-Supervised Domain Adaptation". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8886–8896.

Yin, K., Liu, C., Mostafavi, A., and Hu, X. (June 2024). *CrisisSense-LLM: Instruction Fine-Tuned Large Language Model for Multi-label Social Media Text Classification in Disaster Informatics*.

Yu, Y.-C. and Lin, H.-T. (June 2023). " Semi-Supervised Domain Adaptation with Source Label Adaptation ". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 24100–24109.

Zeng, H., Yue, Z., Shang, L., Zhang, Y., and Wang, D. (Oct. 2024). " Unsupervised Domain Adaptation via Contrastive Adversarial Domain Mixup: A Case Study on COVID-19 ". In: *IEEE Transactions on Emerging Topics in Computing* 12.04, pp. 1105–1116.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "mixup: Beyond Empirical Risk Minimization". In: *International Conference on Learning Representations*.

Zhang, Y., Liu, T., Long, M., and Jordan, M. (2019). "Bridging Theory and Algorithm for Domain Adaptation". In: *International Conference on Machine Learning*, pp. 7404–7413.

Zhou, Z.-H. and Li, M. (2005). "Tri-training: exploiting unlabeled data using three classifiers". In: *IEEE Transactions on Knowledge and Data Engineering* 17.11, pp. 1529–1541.

Method	F1				ECE			
	5	10	25	50	5	10	25	50
California Wildfires 2018								
Qwen3-14B - zero-shot	**0.562**	**0.562**	0.562	0.562	0.217	0.217	0.217	0.217
Supervised	0.424	0.511	0.562	0.626	0.207	0.249	0.201	0.278
Self-Training	0.473	0.545	0.628	0.646	0.256	0.233	0.176	0.164
UDA	0.324	0.410	0.505	0.537	0.198	**0.121**	0.138	0.127
DeCoTa	0.324	0.500	0.611	0.638	0.384	0.178	**0.129**	0.113
DeCoTa - low conf	0.352	0.407	0.529	0.609	**0.179**	0.138	0.143	**0.092**
DeCoTa - weighted conf	0.175	0.425	0.620	0.687	0.610	0.207	0.154	0.097
DeCoTa - cross veiw	0.268	0.488	0.640	**0.693**	0.386	0.192	0.134	0.102
DeCoTa - LLM labeled start	0.433	0.513	**0.641**	0.683	0.221	0.191	0.143	0.099
Cyclone Idai 2019								
Qwen3-14B - zero-shot	**0.605**	**0.605**	0.605	0.605	0.198	0.198	0.198	0.198
Supervised	0.349	0.437	0.539	0.592	**0.190**	0.212	0.321	0.265
Self-Training	0.357	0.469	0.569	0.633	0.260	0.220	0.158	0.141
UDA	0.357	0.396	0.508	0.541	0.300	0.149	0.088	0.176
DeCoTa	0.372	0.514	0.606	**0.664**	0.269	0.146	**0.049**	**0.040**
DeCoTa - low conf	0.333	0.480	0.523	0.539	0.206	0.139	0.177	0.150
DeCoTa - weighted conf	0.175	0.389	0.577	0.662	0.370	0.541	0.180	0.067
DeCoTa - cross veiw	0.431	0.528	**0.613**	0.659	0.267	**0.120**	0.083	0.048
DeCoTa - LLM labeled start	0.282	0.354	0.571	0.642	0.213	0.151	0.110	0.053
Kerala Floods 2018								
Qwen3-14B - zero-shot	**0.474**	**0.474**	0.474	0.474	0.249	0.249	0.249	0.249
Supervised	0.361	0.394	0.457	0.525	0.250	0.188	0.154	0.299
Self-Training	0.354	0.443	**0.568**	0.591	0.292	0.238	0.190	0.159
UDA	0.321	0.383	0.476	0.506	0.236	0.145	**0.100**	0.156
DeCoTa	0.219	0.331	0.540	0.605	0.369	0.304	0.148	0.132
DeCoTa - low conf	0.240	0.352	0.495	0.561	**0.224**	**0.103**	0.106	**0.047**
DeCoTa - weighted conf	0.094	0.271	0.485	0.597	0.594	0.394	0.247	0.116
DeCoTa - cross veiw	0.272	0.384	0.554	0.600	0.395	0.254	0.145	0.143
DeCoTa - LLM labeled start	0.199	0.360	0.528	**0.606**	0.266	0.145	0.123	0.118
Hurricane Dorian 2019								
Qwen3-14B - zero-shot	**0.537**	**0.537**	0.537	0.537	0.292	0.292	0.292	0.292
Supervised	0.404	0.508	0.557	0.577	0.212	0.186	0.334	0.305
Self-Training	0.426	0.524	0.578	0.595	0.361	0.290	0.242	0.231
UDA	0.319	0.400	0.525	0.547	**0.119**	**0.095**	0.168	0.148
DeCoTa	0.271	0.425	0.570	0.597	0.546	0.254	0.204	0.170
DeCoTa - low conf	0.335	0.332	0.393	0.384	0.197	0.131	**0.109**	**0.119**
DeCoTa - weighted conf	0.193	0.407	0.561	0.602	0.589	0.231	0.204	0.176
DeCoTa - cross veiw	0.304	0.421	0.556	0.602	0.484	0.298	0.184	0.185
DeCoTa - LLM labeled start	0.404	0.482	**0.584**	**0.607**	0.247	0.215	0.190	0.188
Hurricane Florence 2018								
Qwen3-14B - zero-shot	**0.637**	**0.637**	0.637	0.637	0.251	0.251	0.251	0.251
Supervised	0.455	0.549	0.637	0.648	0.247	0.222	0.283	0.247
Self-Training	0.509	0.584	0.666	**0.703**	0.302	0.224	0.171	0.139
UDA	0.424	0.524	0.598	0.623	0.176	**0.069**	0.098	0.166
DeCoTa	0.364	0.618	0.661	0.686	0.274	0.173	0.129	0.110
DeCoTa - low conf	0.316	0.588	**0.677**	0.681	0.287	0.103	**0.062**	**0.040**
DeCoTa - weighted conf	0.319	0.481	0.654	0.691	0.285	0.361	0.111	0.086
DeCoTa - cross veiw	0.377	0.616	0.662	0.700	0.386	0.163	0.107	0.123
DeCoTa - LLM labeled start	0.478	0.584	0.671	0.685	**0.167**	0.197	0.111	0.112

Table 2. Cross-domain adaptation performance measured by Macro-F1 and ECE, averaged across target disaster events and multiple source domains for all methods except Supervised and Self-Training. Bold values denote the highest Macro-F1 and the lowest ECE per column. The zero-shot LLM baseline remains constant across all label regimes (5–50), as it does not rely on labeled target data.

APPENDIX

Due to space limitations in the main paper, we provide additional implementation details and extended experimental results in this appendix. Specifically, we include: (i) the complete definitions of evaluation metrics and calibration computation, (ii) detailed hyperparameter for DeCoTa variants including the LLM used for the LLM labeled start, (iii) label-wise performance of the DeCoTa variants on one target disaster, (iv) full per-target numerical results tables (Tables - T1 T2, T3, T4, T5), and (v) supplementary per-target performance plots (Figure S1) averaged across source transfers. These materials provide a comprehensive view of experimental behavior and support the trends summarized in the main text.

Evaluation Metrics

We evaluate performance using **Macro-F1** and **Expected Calibration Error (ECE)** in order to jointly assess classification effectiveness and predictive reliability under domain shift. Macro-F1 accounts for class imbalance across humanitarian categories and ensures equal weighting of minority and majority classes. For each class c, the F1-score is defined as:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{5}$$

where

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}. \tag{6}$$

Macro-F1 is computed as the unweighted mean across the set of target-supported classes C_T:

$$\text{Macro-F1} = \frac{1}{|C_T|} \sum_{c \in C_T} F1_c, \tag{7}$$

where C_T includes only labels with non-zero support in the target test set.

To evaluate model calibration, we compute Expected Calibration Error (ECE), which measures the discrepancy between predicted confidence and empirical accuracy. Predictions are partitioned into M confidence bins. For each bin B_m, let $\text{acc}(B_m)$ denote empirical accuracy and $\text{conf}(B_m)$ the mean predicted confidence. ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{8}$$

where n is the number of evaluation instances. Lower ECE indicates better calibration.

Hyperparameters

All models are optimized using Adam with a learning rate of 5×10^{-5}. A supervised batch size of 16 is used for labeled data, and an unsupervised batch size of 64 is used for methods leveraging pseudo-labeled target instances. Supervised teacher models and all DeCoTa variants are trained for 15 epochs. Self-Training (ST) and UDA-based methods include an additional unsupervised refinement stage consisting of 12 pseudo-label iterations per run. Remaining hyperparameters are selected using validation performance under consistent search ranges to ensure fair comparison. Model calibration is computed using uniform binning with $M = 10$ bins across all methods. Training is conducted on NVIDIA A5000 GPUs.

DeCoTa-Variants:

All DeCoTa variants use the same backbone, where two classifiers are trained with pseudo-label exchange on unlabeled target data. A confidence threshold of $\tau = 0.5$ is used for pseudo-label selection in the base setting, while the low-confidence variant selects samples below τ. The cross-view variant performs pseudo-label exchange across classifiers and applies logit-level MixUp between pseudo-labeled and supervised batches from opposite views. MixUp uses a Beta distribution with $\alpha = 0.4$. In the weighted-confidence variant, pseudo-label losses are scaled by prediction confidence. In the LLM-Start variant, pseudo-labels are sampled between LLM and model predictions for the first $E_{\text{switch}} = 2$ epochs using $s = \max(0, 1 - \frac{e+1}{E_{\text{switch}}})$, where e is the epoch. After this phase, only model-generated pseudo-labels above τ are used.

LLM Configuration:

We use Qwen (Qwen3-14B) for zero-shot pseudo-labeling. Inference uses greedy decoding (temperature = 0) for deterministic outputs. The model is prompted with the tweet text and the label set for the current dataset split. The prompt format is:

> You are a tweet classification assistant. Classify the tweet into exactly one of the given categories below.
>
> Tweet: "<tweet>"
>
> Categories:
> <label id 0>. <label name>
> <label id 1>. <label name>
> ...
>
> Respond in the exact format:
> Answer: <Category number>. <Category text>
> Explanation: <Short explanation>

Label-wise Analysis

Figure S2 presents label-wise F1 across varying supervision levels. Performance improves consistently with increasing labels, with the largest gains observed from 5 to 25 labels per class; however, improvements remain uneven across categories.

Across DeCoTa variants, distinct patterns emerge. LLM-labeled start variant consistently provides strong performance in low-label regimes (5–10 labels), particularly for higher correctly predicted classes by the LLM, such as *sympathy_and_support* and *rescue_volunteering_or_donation_effort*, indicating the benefit of clearer lexical patterns and lower ambiguity. Cross-View variant shows more stable and balanced improvements across mid-difficulty classes (e.g., *infrastructure_and_utility_damage* and *other_relevant_information*), suggesting effective regularization through cross-view interaction. In contrast, Low-Confidence variant tends to underperform across most labels, while Weighted-Confidence variant becomes more competitive at higher label budgets, indicating that confidence-based weighting is more effective when sufficient supervision is available.

Despite these gains, challenging classes such as *not_humanitarian*, *requests_or_urgent_needs*, and *injured_or_dead_people* remain consistently difficult across all variants and label budgets, highlighting persistent ambiguity and data sparsity issues.

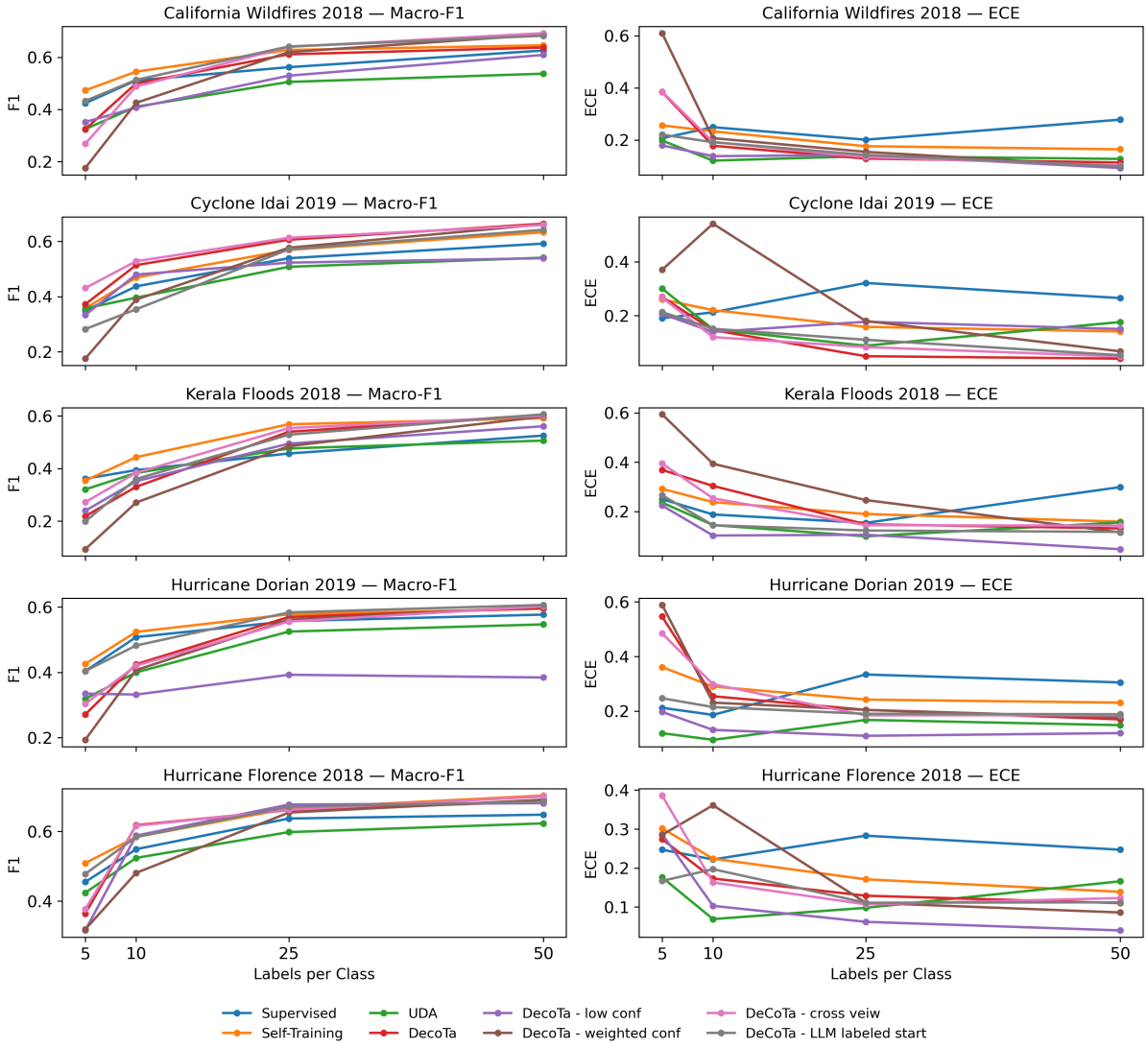

Figure S1. Per-target cross-disaster adaptation results under limited target supervision. Each row shows average Macro-F1 (left) and Expected Calibration Error (ECE) (right) across label regimes, aggregated over all source-to-target transfers for the corresponding target event. DeCoTa and its variants are compared against supervised, self-training, and UDA baselines.

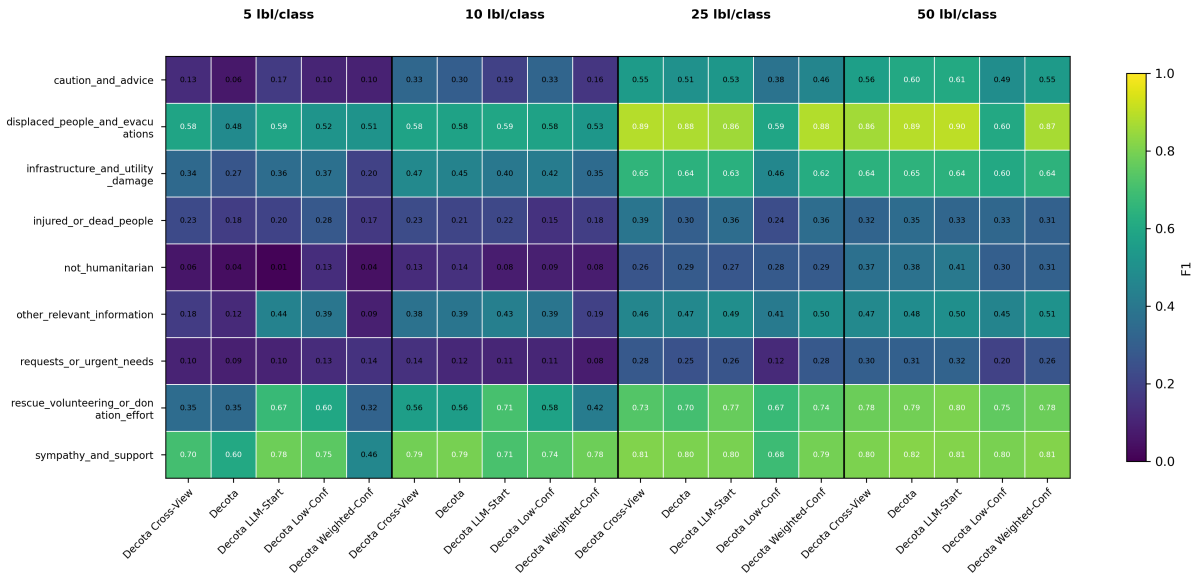

Figure S2. Label-wise F1 performance across DeCoTa variants under varying label budgets (5, 10, 25, 50 labels per class). Results are averaged across three source disasters for the target disaster Hurricane Dorian 2019. Each block corresponds to a label budget, and columns within each block represent different method variants. Performance improves consistently with increasing supervision, with notable variation across classes and methods.

Source	Method	F1				ECE			
		5	10	25	50	5	10	25	50
Baselines									
–	Qwen3-14B - zero-shot	**0.562**	0.562	0.562	0.562	0.217	0.217	0.217	0.217
–	Supervised	0.424	0.511	0.562	0.626	0.207	0.249	0.201	0.278
–	Self-Training	0.473	0.545	0.628	0.646	0.256	0.233	0.176	0.164
Domain adaptation (by source event)									
Canada Wildfires 2016	UDA	0.303	0.353	0.507	0.549	0.142	0.086	0.236	0.232
Canada Wildfires 2016	DeCoTa	0.474	**0.584**	0.444	0.474	0.152	0.189	0.121	0.143
Canada Wildfires 2016	DeCoTa - low conf	0.433	0.384	0.571	0.614	0.194	0.196	0.146	0.107
Canada Wildfires 2016	DeCoTa - weighted conf	0.246	0.412	0.616	0.696	0.481	0.255	0.175	0.123
Canada Wildfires 2016	DeCoTa - cross veiw	0.301	0.516	0.640	0.696	0.420	0.223	0.136	0.105
Canada Wildfires 2016	DeCoTa - LLM labeled start	0.365	0.486	0.654	0.680	0.353	0.262	0.164	0.106
Hurricane Harvey 2017	UDA	0.329	0.453	0.524	0.539	0.235	**0.078**	0.140	0.051
Hurricane Harvey 2017	DeCoTa	0.272	0.525	0.654	0.697	0.406	0.151	0.122	0.077
Hurricane Harvey 2017	DeCoTa - low conf	0.356	0.470	0.565	0.564	0.179	0.082	0.106	0.087
Hurricane Harvey 2017	DeCoTa - weighted conf	0.217	0.537	0.606	0.695	0.565	0.169	0.193	0.071
Hurricane Harvey 2017	DeCoTa - cross veiw	0.318	0.567	0.659	0.696	0.342	0.172	0.087	0.112
Hurricane Harvey 2017	DeCoTa - LLM labeled start	0.517	0.557	0.619	0.674	0.193	0.179	0.144	0.088
Hurricane Irma 2017	UDA	0.338	0.389	0.498	0.533	0.225	0.161	0.107	0.133
Hurricane Irma 2017	DeCoTa	0.447	0.550	0.645	0.675	0.268	0.160	0.136	0.096
Hurricane Irma 2017	DeCoTa - low conf	0.434	0.406	0.528	0.606	**0.102**	0.094	**0.057**	**0.044**
Hurricane Irma 2017	DeCoTa - weighted conf	0.267	0.536	0.622	**0.700**	0.559	0.175	0.151	0.101
Hurricane Irma 2017	DeCoTa - cross veiw	0.421	0.578	0.613	**0.700**	0.174	0.126	0.163	0.102
Hurricane Irma 2017	DeCoTa - LLM labeled start	0.496	0.546	0.636	0.684	0.240	0.147	0.159	0.101
Kerala Floods 2018	UDA	0.316	0.412	0.484	0.534	0.203	0.139	0.146	0.111
Kerala Floods 2018	DeCoTa	0.156	0.305	0.657	0.689	0.616	0.234	0.117	0.112
Kerala Floods 2018	DeCoTa - low conf	0.232	0.338	0.541	0.587	0.217	0.184	0.227	0.127
Kerala Floods 2018	DeCoTa - weighted conf	0.073	0.298	0.628	0.664	0.669	0.223	0.135	0.092
Kerala Floods 2018	DeCoTa - cross veiw	0.152	0.320	0.625	0.693	0.424	0.203	0.161	0.089
Kerala Floods 2018	DeCoTa - LLM labeled start	0.343	0.442	0.655	0.693	0.167	0.173	0.109	0.091
Kaikoura Earthquake 2016	UDA	0.335	0.445	0.514	0.530	0.183	0.140	0.060	0.109
Kaikoura Earthquake 2016	DeCoTa	0.270	0.535	0.657	0.652	0.480	0.154	0.147	0.139
Kaikoura Earthquake 2016	DeCoTa - low conf	0.303	0.437	0.442	0.674	0.203	0.133	0.177	0.096
Kaikoura Earthquake 2016	DeCoTa - weighted conf	0.072	0.344	0.627	0.679	0.776	0.213	0.118	0.101
Kaikoura Earthquake 2016	DeCoTa - cross veiw	0.148	0.459	**0.662**	0.678	0.568	0.237	0.123	0.102
Kaikoura Earthquake 2016	DeCoTa - LLM labeled start	0.442	0.536	0.643	0.682	0.150	0.195	0.137	0.107

Table T1. Per-source Macro-F1 and ECE results for the California Wildfires 2018 target event. Results are grouped into a baseline block and source-specific blocks. Underlined values indicate the best result within each block and column, while bold and underlined values denote the overall best for that column across all blocks.

Source	Method	F1				ECE			
		5	10	25	50	5	10	25	50
Baselines									
–	Qwen3-14B - zero-shot	**0.605**	**0.605**	0.605	0.605	0.198	0.198	0.198	0.198
–	Supervised	0.349	0.437	0.539	0.592	**0.190**	0.212	0.321	0.265
–	Self-Training	0.357	0.469	0.569	0.633	0.260	0.220	0.158	0.141
Domain adaptation (by source event)									
Hurricane Irma 2017	UDA	0.357	0.396	0.508	0.541	0.300	0.149	0.088	0.176
Hurricane Irma 2017	DeCoTa	0.372	0.514	0.606	**0.664**	0.269	0.146	**0.049**	**0.040**
Hurricane Irma 2017	DeCoTa - low conf	0.333	0.480	0.523	0.539	0.206	0.139	0.177	0.150
Hurricane Irma 2017	DeCoTa - weighted conf	0.175	0.389	0.577	0.662	0.370	0.541	0.180	0.067
Hurricane Irma 2017	DeCoTa - cross veiw	0.431	0.528	**0.613**	0.659	0.267	**0.120**	0.083	0.048
Hurricane Irma 2017	DeCoTa - LLM labeled start	0.282	0.354	0.571	0.642	0.213	0.151	0.110	0.053

Table T2. Per-source Macro-F1 and ECE results for the Cyclone Idai 2019 target event. Results are grouped into a baseline block and source-specific blocks. Underlined values indicate the best result within each block and column, while bold and underlined values denote the overall best for that column across all blocks.

Source	Method	F1				ECE			
		5	10	25	50	5	10	25	50
Baselines									
–	Qwen3-14B - zero-shot	**0.474**	0.474	0.474	0.474	0.249	0.249	0.249	0.249
–	Supervised	0.361	0.394	0.457	0.525	0.250	0.188	0.154	0.299
–	Self-Training	0.354	0.443	0.568	0.591	0.292	0.238	0.190	0.159
Domain adaptation (by source event)									
Canada Wildfires 2016	UDA	0.283	0.320	0.402	0.414	0.272	0.159	0.105	0.139
Canada Wildfires 2016	DeCoTa	0.190	0.330	0.551	0.604	0.300	0.179	0.147	0.130
Canada Wildfires 2016	DeCoTa - low conf	0.222	0.344	0.485	0.526	0.166	0.058	0.136	0.051
Canada Wildfires 2016	DeCoTa - weighted conf	0.151	0.322	0.499	0.596	0.347	0.227	0.184	0.113
Canada Wildfires 2016	DeCoTa - cross veiw	0.244	0.353	0.528	0.600	0.247	0.183	0.105	0.140
Canada Wildfires 2016	DeCoTa - LLM labeled start	0.139	0.409	0.510	0.589	0.331	**0.049**	0.148	0.125
Hurricane Harvey 2017	UDA	0.321	0.412	0.513	0.546	0.159	0.163	0.086	0.174
Hurricane Harvey 2017	DeCoTa	0.223	0.366	0.566	**0.617**	0.262	0.218	0.140	0.121
Hurricane Harvey 2017	DeCoTa - low conf	0.217	0.407	0.508	0.577	0.236	0.062	0.130	0.067
Hurricane Harvey 2017	DeCoTa - weighted conf	0.106	0.305	0.547	0.606	0.503	0.197	0.090	0.086
Hurricane Harvey 2017	DeCoTa - cross veiw	0.440	**0.530**	**0.587**	0.600	0.307	0.229	0.206	0.193
Hurricane Harvey 2017	DeCoTa - LLM labeled start	0.319	0.330	0.548	0.606	0.216	0.194	0.083	0.102
Hurricane Irma 2017	UDA	0.314	0.392	0.503	0.537	0.238	0.078	0.108	0.158
Hurricane Irma 2017	DeCoTa	0.368	0.424	0.563	0.609	**0.106**	0.181	0.120	0.120
Hurricane Irma 2017	DeCoTa - low conf	0.270	0.333	0.507	0.580	0.276	0.191	**0.063**	0.038
Hurricane Irma 2017	DeCoTa - weighted conf	0.068	0.272	0.443	0.597	0.667	0.561	0.343	0.124
Hurricane Irma 2017	DeCoTa - cross veiw	0.274	0.403	0.585	0.611	0.336	0.073	0.090	0.095
Hurricane Irma 2017	DeCoTa - LLM labeled start	0.176	0.337	0.548	0.615	0.315	0.239	0.123	0.122
Kaikoura Earthquake 2016	UDA	0.365	0.408	0.487	0.527	0.276	0.180	0.099	0.152
Kaikoura Earthquake 2016	DeCoTa	0.094	0.202	0.478	0.589	0.806	0.637	0.186	0.155
Kaikoura Earthquake 2016	DeCoTa - low conf	0.249	0.323	0.478	0.559	0.219	0.100	0.094	**0.033**
Kaikoura Earthquake 2016	DeCoTa - weighted conf	0.049	0.184	0.451	0.590	0.858	0.589	0.369	0.139
Kaikoura Earthquake 2016	DeCoTa - cross veiw	0.131	0.249	0.514	0.589	0.688	0.529	0.177	0.144
Kaikoura Earthquake 2016	DeCoTa - LLM labeled start	0.161	0.365	0.507	0.613	0.203	0.097	0.138	0.121

Table T3. Per-source Macro-F1 and ECE results for the Kerala Floods 2018 target event. Results are grouped into a baseline block and source-specific blocks. Underlined values indicate the best result within each block and column, while bold and underlined values denote the overall best for that column across all blocks.

Source	Method	F1				ECE			
		5	10	25	50	5	10	25	50
Baselines									
–	Qwen3-14B - zero-shot	**0.537**	0.537	0.537	0.537	0.292	0.292	0.292	0.292
–	Supervised	0.404	0.508	0.557	0.577	0.212	0.186	0.334	0.305
–	Self-Training	0.426	0.524	0.578	0.595	0.361	0.290	0.242	0.231
Domain adaptation (by source event)									
Canada Wildfires 2016	UDA	0.343	0.420	0.511	0.514	0.195	0.093	0.167	0.190
Canada Wildfires 2016	DeCoTa	0.290	0.431	0.557	0.603	0.507	0.220	0.251	0.201
Canada Wildfires 2016	DeCoTa - low conf	0.313	0.459	0.503	0.446	0.262	0.146	0.097	0.149
Canada Wildfires 2016	DeCoTa - weighted conf	0.166	0.399	0.560	**0.631**	0.622	0.291	0.181	0.197
Canada Wildfires 2016	DeCoTa - cross veiw	0.354	0.471	0.541	0.608	0.449	0.264	0.214	0.167
Canada Wildfires 2016	DeCoTa - LLM labeled start	0.425	0.482	**0.591**	0.623	0.234	0.256	0.194	0.205
Hurricane Harvey 2017	UDA	0.334	0.422	0.553	0.567	**0.072**	0.143	0.183	0.111
Hurricane Harvey 2017	DeCoTa	0.408	0.541	0.576	0.599	0.414	0.183	0.184	0.146
Hurricane Harvey 2017	DeCoTa - low conf	0.414	0.183	0.184	0.146	0.082	0.085	**0.092**	**0.096**
Hurricane Harvey 2017	DeCoTa - weighted conf	0.313	0.496	0.582	0.589	0.529	0.214	0.187	0.169
Hurricane Harvey 2017	DeCoTa - cross veiw	0.440	0.530	0.587	0.600	0.307	0.229	0.206	0.193
Hurricane Harvey 2017	DeCoTa - LLM labeled start	0.534	**0.558**	0.581	0.590	0.254	0.218	0.197	0.177
Kerala Floods 2018	UDA	0.280	0.358	0.511	0.560	0.090	**0.048**	0.153	0.144
Kerala Floods 2018	DeCoTa	0.115	0.303	0.576	0.588	0.718	0.360	0.178	0.162
Kerala Floods 2018	DeCoTa - low conf	0.278	0.354	0.491	0.561	0.247	0.162	0.139	0.113
Kerala Floods 2018	DeCoTa - weighted conf	0.101	0.326	0.542	0.585	0.616	0.188	0.245	0.163
Kerala Floods 2018	DeCoTa - cross veiw	0.117	0.261	0.539	0.599	0.697	0.402	0.133	0.195
Kerala Floods 2018	DeCoTa - LLM labeled start	0.253	0.407	0.579	0.607	0.253	0.170	0.179	0.183

Table T4. Per-source Macro-F1 and ECE results for the Hurricane Dorian 2019 target event. Results are grouped into a baseline block and source-specific blocks. Underlined values indicate the best result within each block and column, while bold and underlined values denote the overall best for that column across all blocks.

Source	Method	F1				ECE			
		5	10	25	50	5	10	25	50
Baselines									
–	Qwen3-14B - zero-shot	**0.637**	**0.637**	0.637	0.637	0.251	0.251	0.251	0.251
–	Supervised	0.455	0.549	0.637	0.648	0.247	0.222	0.283	0.247
–	Self-Training	0.509	0.584	0.666	**0.703**	0.302	0.224	0.171	0.139
Domain adaptation (by source event)									
Kaikoura Earthquake 2016	UDA	0.424	0.524	0.598	0.623	0.176	**0.069**	0.098	0.166
Kaikoura Earthquake 2016	DeCoTa	0.364	0.618	0.661	0.686	0.274	0.173	0.129	0.110
Kaikoura Earthquake 2016	DeCoTa - low conf	0.316	0.588	**0.677**	0.681	0.287	0.103	**0.062**	**0.040**
Kaikoura Earthquake 2016	DeCoTa - weighted conf	0.319	0.481	0.654	0.691	0.285	0.361	0.111	0.086
Kaikoura Earthquake 2016	DeCoTa - cross veiw	0.377	0.616	0.662	0.700	0.386	0.163	0.107	0.123
Kaikoura Earthquake 2016	DeCoTa - LLM labeled start	0.478	0.584	0.671	0.685	**0.167**	0.197	0.111	0.112

Table T5. Per-source Macro-F1 and ECE results for the Hurricane Florence 2018 target event. Results are grouped into a baseline block and source-specific blocks. Underlined values indicate the best result within each block and column, while bold and underlined values denote the overall best for that column across all blocks.