CryChime: When Large Language Models Learn to Listen to Distant Cries - A Counterfactual PEFT Framework for Urgent Need Detection in Disaster Social Media

Junhong Cai[1]* JUNHONG.CAI@STUD.UNI-HEIDELBERG.DE
Geng Zhao[1]*† ZHAO@CL.UNI-HEIDELBERG.DE
Jiaxin Li[2]* SHELLEYLIMEER@GMAIL.COM
[1]*Institution of Computational Linguistics, Heidelberg University, Heidelberg, Germany*
[2]*Institution of Mathematics and Computational Science, Huaihua University, Hunan, China*

Editors: Hung-yi Lee and Tongliang Liu

Abstract

In recent years, detecting instantaneously expressed urgent needs or requests in disaster-related posts from disaster-affected social media users has become crucial for disaster response and recovery. To address the gap that the performance of current urgent need detectors based on large language models (LLMs) is below requirements on this task from the domain of disaster response, we propose a novel insight: decomposing and inducting post content expressing disaster-induced urgent needs, into disaster event statements and disaster-induced appeals. The former, widely present and highly coarse-grained homogeneous across disaster-related posts, tends to introduce event-induced model bias leading to false recalls; while the latter, characterized by highly personalized, fine-grained and subjective phrasing, often challenge LLMs to allocate appropriate attentions to the corresponding tokens. In light of this, we propose CryChime, a novel model-agnostic parameter-efficient fine-tuning (PEFT) framework. CryChime represents disaster event statements in a bootstrapping style, and then removes the event-induced bias by orthogonal LoRA-based counterfactual learning. As fine-tuning steps increase, CryChime gradually disentangles the domain knowledge for understanding disaster event statements and disaster-induced appeals in candidate posts, then collaboratively leverage them in performing better urgent need detection. Experimental results show that CryChime can more effectively listen to the distant cries from the disaster-affected users.

Keywords: Disaster Response, Urgent Need Detection, Counterfactual PEFT.

1. Introduction

In recent years, increasingly frequent and severe disaster events pose unprecedented threats to human lives, infrastructure, and assets Palen and Anderson (2016). For instance, the 2011 East Japan earthquake resulted in over 22,000 civilian deaths Norio et al. (2011), while most of its affected areas are still being condemned as uninhabitable nowadays Hikichi et al. (2021). The profound impacts of such disasters highlight the requirements on enhancing disaster-response systems to aid in disaster relief, loss accounting, and other related matters. In this context, managing the explosive influx of information from microblogging social medias becomes crucial for online disaster information coordination Alam and Imran (2018); Imran et al. (2016); Nguyen and Rudra (2022), particularly in perceiving the urgent needs

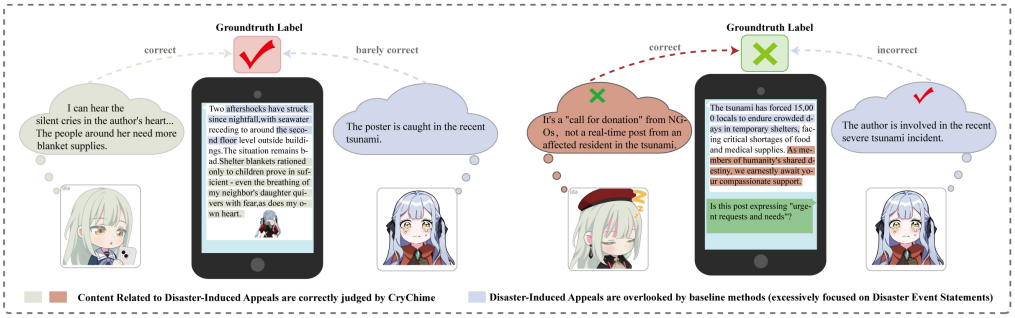

Figure 1: A vivid illustration of the insight of CryChime. Many informative disaster-related posts have a "disaster event statement" (Blue), no matter whether expressing urgent needs or requests. The authors always try to be more objective when writing this part. But the sincere, subjective, personal disaster-induced appeals for their urgent needs/requests (Red/Green) are also crucial for UND.

or requests of those affected by disaster events Lei et al. (2025); Yang et al. (2024b), a pivotal component in disaster response workflows.

Reinforcing the growing importance of urgent need detection mentioned above, the explosive proliferation of SNS since the 2010s has made the task even more critical Reuter et al. (2018), as individuals in disaster-stricken areas can now instantaneously broadcast their urgent needs through personal posts Yang et al. (2024b) on their accounts, providing real-time, first-hand insights regarding their struggles and expectations, to the public. For disaster response coordinators and emergency managers, meeting these urgent needs—whether accurately or by addressing the broader difficulties they reveal—can yield both immediate and high-impact benefits for the post-disaster recovery. Consequently, the urgent need detection in disaster response (We define detecting the individual-level urgent needs or requests expressed in their social media posts during disaster events as the task of this study, abbreviated as urgent need detection (UND).) based on pretrained language models (PLMs) has grasped lots of attention Vitiugin and Purohit (2024); Lamsal et al. (2024) from the research domain of disaster response (on social media).

Since 2023, building on this trend, the growing deployment of large language models (LLMs) has further motivated the exploration Yin et al. (2024) of using LLMs as the backbone for PLM-based urdent need detector—given that LLMs possess both broad world knowledge and strong domain adaptability, and they can effectively follow human instructions—such as focusing on key priorities or adhering to a specific logic—during text understanding, analyzing and generation. These capabilities render LLMs promising for to pursuing the best UND performance. Specifically, the well-pre-trained world knowledge guarantees the robustness of LLMs in understanding disaster-related text and disaster contexts, while instruction-following capability ensures people could request LLMs focus on vulnerable or special groups during disaster events.

∗. Equal Contribution
†. Corresponding Author

However, applying LLMs to UND still faces significant challenges. Specifically, the challenges are generally revolving around that, to combat with the noise induced by data imbalance commonly inherent in disaster NLP tasks Garg et al. (2021); Moreo et al. (2016), how to precisely identify and recall the rare positive samples expressing urgent needs through achieving high-granularity understanding of the posts. According to existing studies Imran et al. (2014); Vitiugin and Purohit (2024); Yang et al. (2024b), the alleviation for these challenges relies on the perspective to deconstruct and inductively analyze the content of positive samples in UND. In light of this, we conduct a qualitative analysis, and through our analysis, we argue that posts expressing disaster-induced urgent needs consist of two main components: relatively objective **disaster event statements** and more subjective **disaster-induced appeals**.

Disaster event statements involve the author's calm, objective description of aspects of the disaster's background (disaster event type), developments, impacts (on the post author) Alam et al. (2021b) . In contrast, disaster-induced appeals encompass both subjective emotional expressions and specific or personal urgent needs related to the post author's current situation and experiencing, often framed as "calls for help." These appeals, although emotionally charged and sometimes lacking in language structure or logic (occasionally dismissed as noise), could convey real urgency and elicit more empathy from the public.

In the real world, a post expressing urgent needs might not present a good cover of the both elements. But from the perspective of human experts, the part of disaster-induced appeals is also non-negligible for the identification. Unfortunately, regarding domain adaptation for LLMs, particularly when the UND fine-tuning dataset is constructed with a social media disaster informatics corpus composed of multiple categories of disaster-related posts, disaster event statements often receive more attention and cause higher probabilistic scores, as they anchor the corresponding posts to the well-observed disaster events in the fine-tuning dataset, which release general warning signs to LLMs. Consequently, LLMs could tend to show higher adaptability to other categories, while sacrificing the performance on UND by neglecting the content of disaster-induced appeals. This will undermine the LLM-based detector's recall capability and robustness. In practical applications, the oversight can lead to the mis-identification of urgent needs as posts about loss of life, donation effort, etc..

To address the challenges, we propose CryChime, a PEFT-based "Chime" for improving urgent need detection by listening to the "cries" during disaster events. Given an LLM-based detector, in the fine-tuning stage, CryChime gradually disentangles and then collaboratively leverages the knowledge respectively learned by analyzing disaster event statements and disaster-induced appeals in candidate posts.

Specifically, to softly represent the disaster event statement from the posts, CryChime employs a Q-Former-based BLIP-2 architecture to learn how to extract and compress disaster event statements into structured soft prompts from general social media disaster NLP datasets. The represented disaster event statements include three aspects: Disaster Event Type, Affected Event Mention and Time Window. We construct question-answering instances for every aspect to pretrain the Q-Former. The combined dataset is named as disaster event statement dataset (DES).

Next, we iteratively conduct two fine-tuning strategies: (a) We combine the UND training datasets and a self-constructed disaster-induced appeals analysis dataset (DIAA). DIAA includes three aspects of instances: Need Type Classification/Extraction, Self-Capacity As-

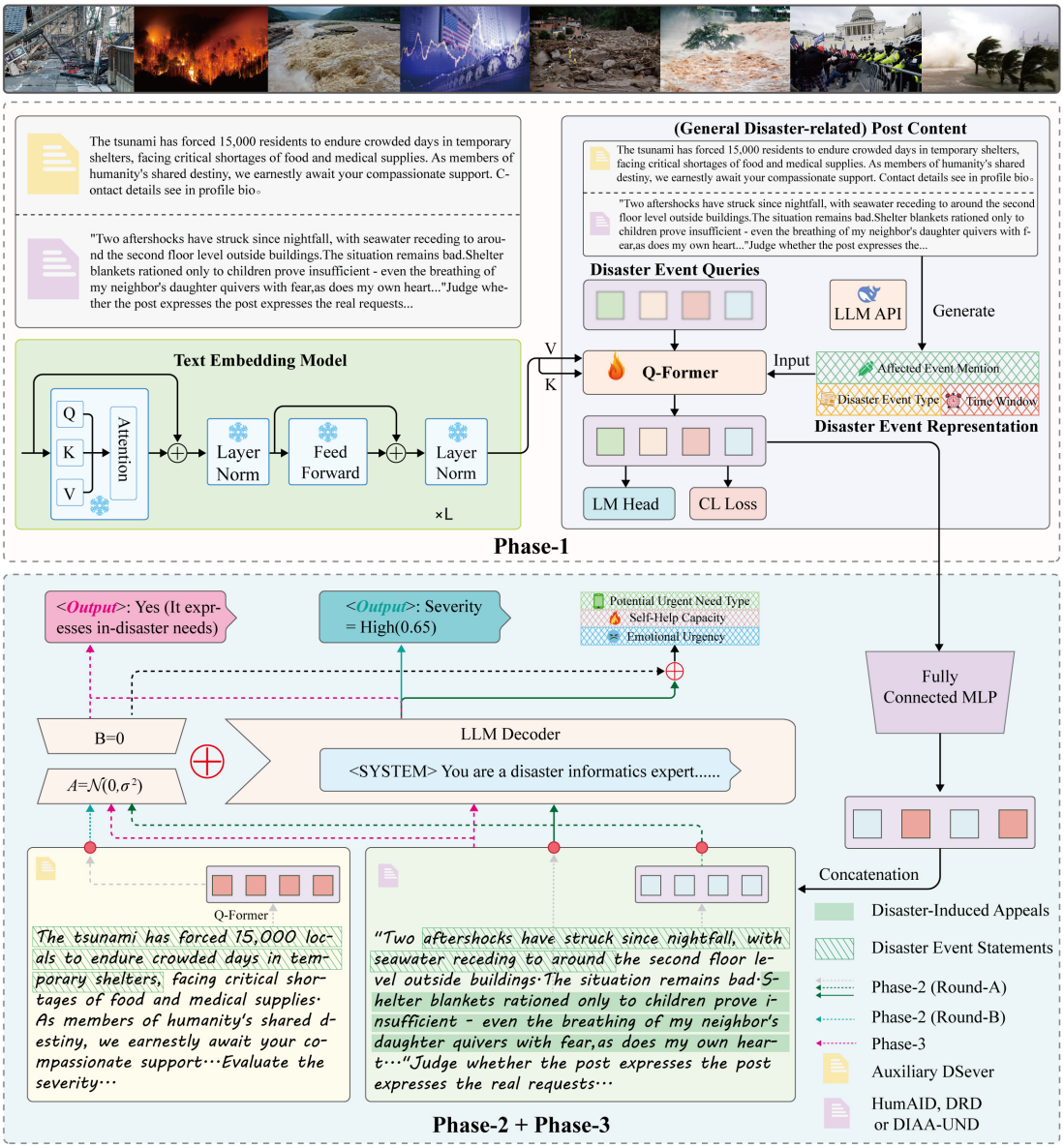

Figure 2: An architecture overview of the proposed CryChime (3-round PEFT workflow). Arrows (dark green) denote the forward passing in Phase-2(Round-A). Arrows (black) denote the forward passing related to Q-Former and soft prompts.

sessment, Emotional Urgency Analysis. The combined dataset is named as DIAA-UND. CryChime uses the extracted soft prompts as counterfactual worlds for counterfactual learning on DIAA-UND, thereby encouraging the model backbone to focus more on disaster-induced appeals alternatively. (b) We perform LoRA-based orthogonal PEFT on an auxiliary dataset (DSever) re-annotated by us, where the labels denote the severity score only

according to the disaster event statements in the corresponding posts, to consolidate knowledge related to disaster event statements in LoRA blocks, and the detector's capability regarding making disaster-related assessment, while mitigating underfitting of the counterfactual causal path. Finally, in the last phase, we conduct instruction-tuning using UND dataset only, achieving collaboration of the debiased backbone (focused on disaster-induced appeals) and the LoRA block.

Experiments on two benchmark datasets of UND show that CryChime outperforms baselines in its capability to detect posts expressing urgent needs or requests during disaster events, with stronger robustness in the face of domain shifts on disaster events.

2. Related Work

As a subfield of disaster response, UND aims to detect emergent requirements of affected populations during disaster events through disaster-related data from social medias. It can directly inform resource allocation and trigger out efficiency-sensitive relief actions.

So far, multiple benchmark datasets systematically curate social media posts expressing urgent needs Imran et al. (2014); Alam et al. (2021a); Imran et al. (2016), and annotate them into a distinct category of disaster-related content Alam et al. (2021a,b). Leveraging these resources, prior studies study UND through multiple insights. Detailedly, some recent studies prompt and fine-tune encoder-only PLMs towards UND Yang et al. (2024b) and additionally give a special focus on rescue-related needs Toraman et al. (2023). Teacher-student model Vitiugin and Purohit (2024) is also introduced to rank the urgency of needs, to improve the robustness and the truthworthiness of detection results. Additionally, a recent study Lamsal et al. (2024) matches urgent needs with resources through similarity-based retrieval methods to optimize resource allocation.

Unfortunately, compared to other subtasks in disaster response, UND is underexplored Lei et al. (2025) and more difficult even for GPT-4o Imran et al. (2025), as its accuracy is below the averaged accuracy across detection subtasks by 15 percent points. To date, UND still faces many addressed challenges including mining voices from real disaster-affected users amid the extremely wide propagation of disaster events, achieving high-granularity capturing and understanding of their emotional expressions and appeals, etc. Some of them are significantly mitigated by our proposed CryChime.

3. Methodology

The framework of CryChime is illustrated in Figure 2. CryChime includes three phases:

Phase-1: We begin by constructing DES using posts from general disaster response classification datasets, then introducing Q-Former to learn to query and compress each post into a soft prompt (composed of multiple learnable tokens) that expresses the three aspects of disaster event statement.

Phase-2: We alternately conduct counterfactual fine-tuning and orthogonal LoRA tuning respectively on DIAA-UND, and on the auxiliary dataset DSever. In this phase, there're two rounds. For each instance, in round-A, by counterfactual learning, we leverage the inferred soft prompt along with the text input to force the detector to learn the domain knowledge for capturing and analyzing disaster-induced appeals. While in round-B, we

#Prompt (Round-B)
[Disaster Event Statement Representations] is a summarized disaster event statement. Assess the threat and urgency. A short answer is: <*Disaster Event Type (1)*>: Which of the following big disaster events is currently happening? Options: earthquake, hurricane, flood, wildfire, cyclone? <*Affected Event Mention (2)*>: Extract the specific disaster-induced event that the post describes. <*Time Window (3)*>: Determine the life stage of the disaster at the time of the text was posted.

#Prompt (Round-A)
[Disaster Event Statement Representations] is a summarized disaster event statement. Full post information: {***Text_Content***}. <*Potential Need Type Estimation (1)*>: Determine the PRIMARY urgent need expressed in the post from categories: Rescue, Medical, Food/Water, Shelter, Protection, Other. <*Self-Help Capacity Evaluation (2)*>: evaluate the poster's self-help capacity (e.g., High/Medium/Low/None), with a less-than-10-word elaboration. <*Emotional Urgency Analysis (3)*>: What is the extent of the author's emotional urgency expressed in the given text? <*UND (4)*>: Is this post expressing urgent need or request?

#Prompt [Phase-3]
[Disaster Event Statement Representations]is a summarized disaster event statement. Assess the threat and urgency. Full post information: {***Text_Content***}. Is this post expressing personal urgent need or request? Answer:

Figure 3: Essentials of Prompts for CryChime (Each aspect has multiple prompt versions).

use the soft prompts only to learn to leverage extracted disaster event statements to make disaster-related assessment, thereby enhancing the negative causal path for the counterfactual world in round-A. The disentanglement of the two kinds of domain knowledge is guaranteed in this stage.

Phase-3: Instruction-tuning with a smaller update rate is conducted dedicatedly on UND to enable the collaboration of the bifurcated domain knowledge acquired in Phase-2.

3.1. Phase-1: Soft In-Context Disaster Event Representation

As shown in Phase-1, Figure 2, we introduce Q-Former from vanilla Blip-2 Li et al. (2023). Initially, we sampled posts from general disaster response datasets as the post content. Then, we construct prompt DES dataset according to the three aspects mentioned above: (1) Disaster Event Type: Determine the macroscopic disaster event which causes the given post (Classification). (2) Affected Event Mention: From the post content, extract a textual mention of a fine-grained, disaster-induced event that has directly impacted the post author or their immediate environment. (3) Time Window: Determine the life stage of the disaster at the time of the text was posted (Classification or Question Answering). Next, we combine the prompts and the post content, call a SOTA LLM API[1] to generate the silver answer as the field "output" of data instances in DES. For each instance, the corresponding silver answer is encoded, then serves as the text input of the Q-Former (initialized with BERT's weights) for its pretraining.

Subsequently, we initialize 16 query tokens for Q-Former. In Q-Former, the query tokens query the text input by cross-attention. Then, the transformed query tokens and text input representations are respectively output by the Q-Former. For pre-training, we calculate a contrastive alignment loss and a text generation loss using the output query tokens and the output text input representations (aligned with the BLIP2's original ITC and ITG respectively). Finally, we conduct inference and generate soft prompts for all samples used in Phase-2 and Phase 3. These soft prompts are exactly the representations of disaster event statements.

1. https://platform.deepseek.com/, we apply DeepSeek-R1

Table 1: Dataset statistics of HumAID Alam et al. (2021a) (restrictive open-sourced)

	HumAID Alam et al. (2021a)	
Partition	Urgent Needs	Total Posts
2016 Earthquakes (Ecuador; Italy; Kaikoura)	145	5051
2017 Hurricanes (Harvey; Maris; Irma)	1170	25959
2017 Sri Lanka Floods	34	575
2017 Mexico Earthquake	61	2036
2018-2019 Test Set (8 Real-World Disaster Events)	1136	39657

3.2. Phase-2 (Round-B): Orthogonal LoRA for Disaster Severity Assessment

From the DES dataset used in Q-Former pretraining, and the UND training set , we sample a subset of their post content. For each post, we annotate a severity score reflecting the threat and urgency level of the disaster event statement in it, thereby creating an auxiliary dataset named as DSever. Subsequently, we insert a LoRA Hu et al. block into the frozen LLM, use Q-Former-generated soft prompts as the primary inputs, to conduct LoRA-based fine-tuning on DSever, aiming to boost the model's capability to assessing the severity of disaster-related posts. By improving the alignment between this domain-specific knowledge and the UND task, this approach enhances the model's capacity to effectively utilize the extracted disaster event statements. Considering the instability of knowledge representations produced by the counterfactual fine-tuning (see in Section 3.3), and to prevent potential knowledge conflicts (which might cause catastrophic forgetting) between the LoRA (counterfactual branch) and the backbone (main branch; focusing more on disaster-induced appeals), we approximate a low-rank SVD decomposition of the frozen (fine-tuned) backbone weights, then construct a regularization term between the LoRA and the decomposition results, to ensure the orthogonality between their feature subspaces, thus alleviating the above concerns. It can be written as:

$$\mathcal{L}_{orth} = \sum_{i_1,\ i_2} ||(\sqrt{\hat{\Sigma}}\hat{V}^T A)[i_1,\ i_2]||^2, \tag{1}$$

where $(\sqrt{\hat{\Sigma}}\hat{V}^T)$ denotes the low-rank reconstruction of the decomposed backbone weights in the format of $A \in \mathbb{R}^{r\times d}$. $\hat{V}^T \in \mathbb{R}^{r\times d}$ is calculated by $\text{SVD}_{\text{low-rank}}(W) = \hat{U}\hat{\Sigma}\hat{V}^T$, where W denotes the frozen pre-trained weights of LLM backbone (epoch $= 0$) or the weights from the latest counterfactual fine-tuning round (epoch > 0). According to the empirical finding in O-LoRA Wang et al. (2023), B is exempted from regularization.

3.3. Phase-2 (Round-A): Bootstrapping LLM to focus on Disaster-Induced Appeals by Counterfactual Learning

In this part, we aim to guide the neurons in the LLM-based detector's backbone to more sensitively and deeply perceive and understand disaster-induced appeals, as they encompass the most sincere, instant emotional expressions and calls for help from those affected by the disaster events. It's immensely valuable both from the standpoint of enhancing UND performance and improving humanitarian assurances. In light of this, we conceptualize the

"counterfactual world" as one where the model, in an indifferent manner, solely relies on the disaster event statements to mechanically determine the presence, validity and other relative aspects of a potential request or need. Thus, the "soft-prompt-only" is considered a counterfactual input as it blocks out any disaster-induced appeals.

Another sense of the Round-B for the Round-A lies in that the LoRA, trained in previous alternated training iterations, acts as a strong projection which models the correlation (direct causal effect) between the counterfactual input and the need assessment results. In previous studies Lu et al. (2024); Zhu et al. (2022), the placement of the fusion between the counterfactual model branch and the backbone, typically at the downstream or task-specific head, tends to impair its influence on the gradient update direction of the model backbone. In contrast, our insight of treating the LoRA blocks as the counterfactual branch, allow the counterfactual model branch to be in parallel with the model backbone throughout. The counterfactual learning causal graph of CryChime is shown in Figure 4

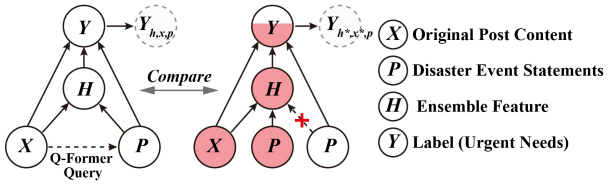

Figure 4: Causal graph of CryChime ("H" mainly denotes disaster-induced appeals).

In detail, the fine-tuning dataset used in Round-A is a combination (DIAA-UND) of 20% of the UND training datasets and a disaster-induced appeals analysis dataset (DIAA) constructed by us.

The task prompts in DIAA cover three aspects related to urgent need expressions: (1) Potential Need Type: Determine (the primary) urgent need expressed in the post from categories: Rescue, Medical, Food/Water, Shelter, Protection, Other. (2) Self-Help Capacity: Evaluate the author's self-help capacity and briefly analyze the author's vulnerability (3) Emotional Urgency: Evaluate the extent of the author's emotional urgency expressed in the text.

Next, for the factual world, we apply zero-masking to each token in the soft prompt with a 50% probability, then concatenate it with the post content to form an factual input, which is:

$$\mathcal{O}_h = f_h(m(x \oplus \tilde{h}_p)), \tag{2}$$

where $m(\cdot)$ denotes the LoRA-plugged-in LLM, $\tilde{h}_p$ denotes the soft prompt after masking operation, and $\oplus$ is a concatenating operation. The LoRA block learned in Round-A is unlearnable for $\mathcal{O}_h$. Parallel to this, for the counterfactual world, we feed the model with unmasked soft-prompt-only as the counterfactual input, which is:

$$\mathcal{O}_p = f_p(m(h_p)), \tag{3}$$

where the formula omits a detach operation exerted on $m(h_p)$. $f_h(\cdot)$ and $f_h(\cdot)$ respectively denote the task heads. Finally, the counterfactual fine-tuning can be written as:

$$\mathcal{L} = \mathcal{L}([(1-\alpha)\mathcal{O}_h + \alpha\mathcal{O}_p], y) + \beta\mathcal{L}(\mathcal{O}_p, y), \tag{4}$$

where $[(1-\alpha)\mathcal{O}_h + \alpha\mathcal{O}_p]$ is the Total Effect (i.e., $Y_{h,x,p}$) shown in figure 4, and $\mathcal{O}_h$ represents the Total Indirect Effect (i.e., $Y_{h,x,p} - Y_{h^*,x^*,p}$), which is the disaster-induced appeals-aware detection of urgent needs given by CryChime.

Empirically, as the counterfactual model is required to explicitly model the causal path independent from the factual world, it's often configured as a more lightweighting branch compared to the base model. Otherwise, the computational expenditure could be excessive. On the contrary, the underfitting of the counterfactual branch could also negatively impact the debiasing. While CryChime, however, allows counterfactual inputs to also leverage the pre-trained LLM weights, without introducing too many additional parameters. Furthermore, we emphasize that although an excessive focus on disaster event statements is more detrimental than beneficial, a moderate focus does not. Thus, for UND, the disentanglement-integration strategy (to balance LLM's attention distribution on the two parts) used by CryChime is better than the "complete removal" strategy.

Consequently, our alternative counterfactual learning strategy allows for the potential collaboration of knowledge from the well-trained, debiased factual world and the counterfactual world in a later stage (see in Para 3.4). That's because the domain knowledge regarding disaster event statements is not completely excluded as confounder but is instead filtered out and stored in the orthogonal LoRA, which can later be further utilized for collaboration with the fine-tuned model backbone that focuses more on disaster-induced appeals, thereby achieving better UND performance.

3.4. Phase-3: Last Fine-Tuning for Collaborative Urgent Need Detection

In this phase, through simply conducting post-instruction-tuning on existing UND training datasets, we encourage CryChime to primarily focus on disaster-induced appeals in user posts, in order to extract the poster's in-disaster urgent needs. Additionally, CryChime can also anchor the background information in the posts to the associated disaster events and evaluate them as a support to the former. In the workflow, the two insights are implemented respectively through the LLM's backbone and the orthogonal LoRA. Meanwhile, because of domain knowledge for mining appeals from subjective in-disaster narratives and assessing the urgency within disaster event statements are learned independently, Phase-3 could also be regarded as arousing the integration and collaboration between them. More illustrations about the insights of this part see in Section 3.3.

4. Experiments

4.1. Experiment Settings

4.1.1. Datasets, Task Setups and Implementation Details.

We compare our proposed CryChime against baseline methods on two of the most authoritative benchmark UND datasets, namely HumAID and DRD.

HumAID Alam et al. (2021a): HumAID comprises approximately 77,000 tweets extracted from 19 disaster events (clearly labeled) that occurred from 2016 to 2019. It covers a variety of disaster event types and is for training models capable of understanding disaster-related information. In HumAID, posts officially labeled with "request or urgent needs" are considered as positive samples for UND. We construct the training set by selecting disaster

Table 2: Comprehensive evaluation on Hum-AID and DRD Datasets. The best results that pass $p \leq 0.005$ paired t-test are shaded. Metrics: Recall ± Std; Precision ± Std

		LLaMA-3.1-8B		**LLaMA-3.2-3B**		**Qwen-2.5-14B**	
Dataset	**Method**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
HumAID	ZS+CoT	65.42	44.14	63.11	42.43	66.41	51.06
	SFT	80.15	75.79	81.69	80.68	84.71	81.07
	SFT+LoRA	81.84	80.59	82.55	82.20	84.30	83.71
	O-LoRA	83.31	77.25	80.68	75.31	81.45	79.23
	ME-MoE	83.62	83.02	83.94	83.24	83.13	84.24
	CryChime	**85.27**	**88.69**	**85.32**	**89.43**	**86.96**	**91.02**
DRD	ZS+CoT	66.28	54.73	67.69	52.55	72.41	56.84
	SFT	85.03	85.95	85.91	87.08	84.79	91.06
	SFT+LoRA	85.92	87.47	85.15	88.26	85.22	91.51
	O-LoRA	86.53	82.39	86.73	83.32	87.75	85.17
	ME-MoE	85.10	88.31	86.99	87.94	86.82	91.84
	CryChime	**87.59**	**93.55**	87.32	**93.06**	87.59	**95.19**

events in 2016-2017, where the proportion of positive samples of each selected events must exceed 1%, and we build the test set with all events in 2018-2019. HumAID contains a larger number of posts categorized under causalty reports, donation efforts, and sympathy support, which are highly possible to result in false-positive predictions. Data statistics is shown in Table 1.

DRD[2]: DRD (Disaster Response Dataset) consists of tweets collected during various disaster events that occurred in 2010 and 2012. We used the re-annotated version provided by CrisisBench Alam et al. (2021b).

For our constructed datasets, i.e. DES (Section 3.1), DSever (Section 3.2) and DIAA (Section 3.3), details are:

DSevers: We construct DES as the pre-training dataset for the Q-Former's learning of how to query and compress disaster event statements into soft prompts. We extract 50,000 samples from the general disaster response datasets provided by CrisisBench and remove all samples categorized as "request or urgent needs" Then convert them into instances by prompting DeepSeek-R1 to generated the output with the post content and the three aspects of task prompts. We also synthesize 10% of the randomly selected UND training data into DES instances.

DSever: We merge the post content of a subset of the DES dataset (with a 50% random probability) with 40% of the UND training set to form the corpus for Round-B. Human annotators assign a 0-1 severity score to each sample's disaster event statements.

DIAA: Instances of DIAA are synthesized following the LLM Api prompting workflow for synthesizing DES. The key difference is, for each aspect of disaster-induced appeals (Poten-

2. https://www.appen.com/datasets/combined-disaster-response-data

 # Task Instruction
You are an expert in disaster informatics. You will receive a post from a user who is possibly affected by a disaster event. You must:

#CoT
(1): Assess the threat and urgency of the disaster event in the post,
(2): Use analytical reasoning to step by step deduce the sufferings and the difficulties of the user.
(3): Last, consider your assessment(1) and reasoning(2) results, answer "Yes" or "No" to the question: is this post expressing personal urgent need or request?

 # User Input
Post Content: {***Text Content***}. Your answer is:

Figure 5: Prompt for baseline "ZS+CoT" and Section 4.3 "LLM-incorporated Evaluation".

tial Need Type, Self-Help Capacity and Emotional Urgency), considering data imbalance in UND datasets, we only generated 3000 instances (9000 in total). 30% of them are from UND dataset's positive samples.

As previous studies Kostina et al. (2025) indicates that input template with task instruction and CoT-included prompt can enhance the model performance. Therefore, we select the chat version of LLaMA-3.2-3B, LLaMA-3.1-8B and Qwen-2.5-14B Yang et al. (2024a) as the base models.

In Phase-2 (Round-B), we additionally incorporate $||W - \hat{U}\hat{\Sigma}\hat{V}^T||^2$ as a regularizer to mitigate the error caused by low-rank approximate SVD decomposition. In Phase-2 (Round-B), since counterfactual finetuning relies on probabilistic outputs, we deploys a classification head. In Phase-2, Round-A and Round-B alternate one time in each epoch, with Round-B being conducted first. In our experiments, we run the experiments on 4 NVIDIA A100 80G GPUs. For CryChime, We conduct Phase-1 Q-Former pre-training, Phase-2 (round-A and round-B alternatively in one epoch) and phase-3, respectively for 3/4/4 epochs.

4.1.2. Baselines

In this paper, we select 5 baselines as follows:
ZS+CoT: In line with the insights conveyed CryChime, we draft an instruction template (sees in Fig 5) with CoT to perform zero-shot (ZS) prompting on the base LLMs.
SFT: Supervised fine-tuning on UND.
SFT+LoRA: After finishing SFT, we integrate LoRA blocks into the model to perform an additional LoRA-based PEFT. The epoch number is set equal to that of SFT.
O-LoRA Wang et al. (2023): Orthogonalized LoRA for continual PEFT. In our experiments, we partition all samples into multiple sub-datasets based on their associated disaster event types. Each sub-dataset is treated as a subtask for O-LoRA, and a corresponding set of LoRA weights is trained. Subsequently, the LoRA weights from all groups are aggregated to form the initial model update.
ME-MoE Nan et al. (2021); Ying et al. (2023): Adding an event-type-specific MoE (Mixture of Experts) module downstream of the LLM to mitigate the negative impact of disaster event statements. Specifically, we create multiple lightweight experts downstream of the LLM. Each expert is initialized as a 1D attention layer + linear layer. The top-1 hard router exclusively receives the soft prompt, reflecting that routing decisions are made based on the disaster event statementn.

Table 3: Prompt-based evaluation on Hum-AID and DRD Datasets. Api-M1: GPT-4o; Api-M2: DeepSeek-V3. "Base-Ori" denotes the zero-shot prompting version w/o the CoT derived from CryChime's insights, but only simply using "Let's think step by step" instead. The best results that pass $p \leq 0.005$ paired t-test are shaded. Metrics: Recall ± Std; Precision ± Std

Dataset	**Metric**	**M1-ZS**	**M1-FS**	**M2-ZS**	**M2-FS**	**M1-Ori**	**M2-Ori**	**Ours**
HumAID	P	78.06	77.95	82.41	79.84	77.84	80.25	87.91
	R	83.53	82.80	80.05	78.34	79.26	76.79	92.26
DRD	P	82.35	82.92	84.1	82.26	82.86	83.19	88.20
	R	89.13	87.27	86.37	85.19	85.53	83.15	95.42

For both SFT and SFT+LoRA, we experiment with either concatenating a soft prompt to the input or not and report the better.

4.2. Comprehensive Evaluations

Comprehensive evaluation across datasets and baselines are reported in Table 2. In real-world applications filled with noise, the proportion of urgent need expressions is quite low. Thus, detecting and recalling positive samples becomes the primary challenge. Furthermore, consider the recalled posts often need to be evaluated by human experts, the proportion of true positives is also important for the efficiency. Thus, we report Precision and Recall scores. Our findings are as follows:
(a): CryChime outperforms the baselines across all experimental groups. Compared to the best result of the baseline methods, it shows an advantage of 3.20 percentage points (pp).
(b): As a benchmark dataset more representative of real disaster scenarios, with data characteristics like temporal OOD (Out-of-Domain), HumAID poses a greater challenge for the models in deeply mining and recalling urgent needs or requests. Excitingly, CryChime demonstrates an exceptional Recall score on HumAID, which is on average 6.21 pp higher. This suggests that the strategy of disentangling then integrating regarding the LLM's knowledge for disaster event statements and disaster-induced need appeals, leads to impressive performance lift in recalling the rarely appeared urgent need or request expressions. CryChime also outperforms the baseline by 4.46 pp on DRD, thereby exhibiting its advantage on standard scenarios.

4.3. LLM-incorporated Evaluation

In this part, considering the concerns about local training costs and inference efficiency, CryChime is built upon small LLMs. Meanwhile, in resource-constrained scenarios, applying prompt engineering to SOTA LLMs is also a widely adopted approach. Therefore, in addition to the baselines, we compare CryChime with SOTA LLMs enhanced by prompt engineering. Specifically, following existing works, we write the insight of CryChime (disaster event statements + disaster-induced appeals) as CoT of the prompts. We select GPT-4o

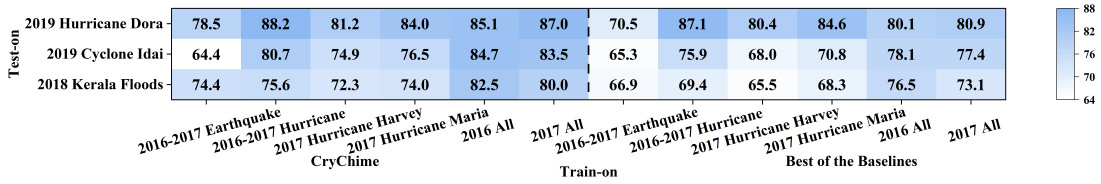

Figure 6: Results of disaster event transfer evaluation (Model: Qwen-2.5-14B, Metric: F1).

and DeepSeek-v3 for the experiment and conduct the evaluation in two settings: 3-instance few-shot (FS) and zero-shot (ZS). Details of the prompt for LLM APIs see in Figure 5.

As shown in Table 3, CryChime(Fine-tuned on the both UND training sets) outperforms "SOTA LLM + prompt engineering" by 6.16 pp, indicating that CryChime, as a specialized small LLM, can serve as an alternative to LLM APIs. Furthermore, the prompt based on the insights of CryChime can lift the need identification ability of SOTA LLMs more close to that of the expert models after SFT.

4.4. Ablation Study

We construct four degraded versions of CryChime in order to better understand the contribution of each component within the full framework. Specifically, we gradually remove different modules from the complete model, and each modified variant is regarded as a degraded baseline.

Concretely, the four degraded versions are obtained as follows: Dg-1 is produced by removing the Q-Former module, Dg-2 is generated by discarding the counterfactual learning strategy, Dg-3 excludes the orthogonal LoRA PEFT mechanism, and Dg-4 simplifies the input into a soft-prompt-only form. These variants are denoted Dg-1 through Dg-4, respectively. More algorithmic details and implementation specifications can be found in the Supplementary Materials.

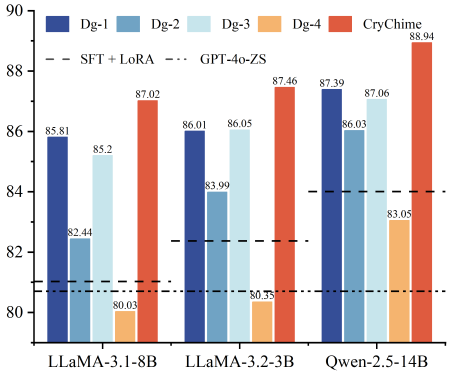

Figure 7: Results of ablation study (Backbone: Qwen-2.5-14B, Metric: F1-Score).

The results show that soft prompts generated by Q-Former is of higher quality than textual extracted disaster-event representations (Dg-1). However, the soft prompt is clearly insufficient to fully and deeply express urgent needs or requests (Dg-4). The fine-grained perception of disaster-induced appeals for urgent needs and knowledge disentanglement via counterfactual learning in the insights indeedly largely contributed to UND(Dg-2). Additionally, orthogonal LoRA tuned on DSever contributes to the performance gain regarding recall capability (Dg-3).

4.5. Domain Transfer Evaluation

To evaluate CryChime's transfer ability across both time and events (i.e., its cross-domain and temporal robustness), we sample from the disaster event collection and timeline from the training set of HumAID, to create six sub-datasets for training. Subsequently, we select three disaster events with sufficient positive samples—2018 Kerala Floods, 2019 Cyclone Idai, and 2019 Hurricane Doran—as test subsets.

The experimental results shown in Figure 6 indicate that, beyond its general advantages, in cross-event groups (e.g., training on 2016–2017 Earthquake data and testing on 2018 Kerala Floods), CryChime's advantage increases to an average of approximate 7.1 pp, demonstrating its strong cross-domain robustness. Moreover, transferring over time does not cause obvious impact on the performance gap.

5. Conclusions

We propose CryChime, a PEFT-based framework that disentangles disaster event statements and disaster-induced appeals to improve the detection of affected people's urgent needs and requests in disaster response. Through experiments and illustrations, we successfully demonstrate that CryChime offers a reliable help for real-world disaster response systems to hear more distant cries during disaster events.

Acknowledgements

This work has been mentored by Dr. Daniel Dahlmeier from SAP, SE. We deeply thank his great mentorship. We also thank the technical assistance provided by Mr. Xiaoci ZHANG and Ms. Yuzhou. They are both second-year bachelor students at Department of Computational Linguistics, Heidelberg University.

References

Firoj Alam and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

Firoj Alam, Umair Qazi, and Ferda Ofli. 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, 2021a.

Firoj Alam, Hassan Sajjad, and Ferda Ofli. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932, 2021b.

Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. Towards robustness to label noise in text classification via noise modeling. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3024–3028, 2021.

Hiroyuki Hikichi, Jun Aida, Katsunori Kondo, and Ichiro Kawachi. Six-year follow-up study of residential displacement and health outcomes following the 2011 japan earthquake and tsunami. *Proceedings of the National Academy of Sciences*, 118(2):e2014226118, 2021.

Edward J Hu, Phillip Wallis, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd international conference on world wide web*, pages 159–162, 2014.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 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, 2016.

Muhammad Imran, Abdul Wahab Ziaullah, Kai Chen, and Ferda Ofli. Evaluating robustness of llms on crisis-related microblogs across events, information types, and linguistic features. In *Proceedings of the ACM on Web Conference 2025*, pages 5117–5126, 2025.

Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*, 2025.

Rabindra Lamsal, MariaRodriguez Read, Shanika Karunasekera, and Muhammad Imran. Crema: Crisis response through computational identification and matching of cross-lingual requests and offers shared on social media. *IEEE Transactions on Computational Social Systems*, 2024.

Zhenyu Lei, Yushun Dong, Weiyu Li, and Jundong Li. Harnessing large language models for disaster management: A survey. *arXiv preprint arXiv:2501.06932*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Junyu Lu, Bo Xu, Xiaokun Zhang, Kaiyuan Liu, Dongyu Zhang, Liang Yang, and Hongfei Lin. Take its essence, discard its dross! debiasing for toxic language detection via counterfactual causal effect. In *LREC-COLING 2024*, pages 15566–15578, 2024.

Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016.

Qiong Nan, Juan Cao, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347, 2021.

Thi Huyen Nguyen and Koustav Rudra. Rationale aware contrastive learning based approach to classify and summarize crisis-related microblogs. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 2022.

Okada Norio, Tao Ye, Yoshio Kajitani, Peijun Shi, and Hirokazu Tatano. The 2011 eastern japan great earthquake disaster: Overview and comments. *International Journal of Disaster Risk Science*, 2:34–42, 2011.

Leysia Palen and Kenneth M Anderson. Crisis informatics—new data for extraordinary times. *Science*, 353(6296):224–225, 2016.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, 34(4):280–294, 2018.

Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Umitcan Sahin. Tweets under the rubble: Detection of messages calling for help in earthquake disaster. *arXiv preprint arXiv:2302.13403*, 2023.

Fedor Vitiugin and Hemant Purohit. Multilingual serviceability model for detecting and ranking help requests on social media during disasters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1571–1584, 2024.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Pingjing Yang, Ly Dinh, Alex Stratton, and Jana Diesner. Detection and categorization of needs during crises based on twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1713–1726, 2024b.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*, 2024.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, and Shiming Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 5384–5392, 2023.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125, 2022.